ARTIFICIAL IDIOCY

Come l'Intelligenza Artificiale è diventata Stregoneria Digitale

Alessandro Parisi

Innovation eXploited

ARTIFICIAL IDIOCY
COME L'INTELLIGENZA ARTIFICIALE È DIVENTATA STREGONERIA DIGITALE

Alessandro Parisi

PROLOGO

“La Verità rende Liberi”

INTRODUZIONE

Parlare di Intelligenza Artificiale é ormai di moda: del resto, le innovazioni che caratterizzano il settore sono talmente rapide ed esaltanti che é difficile resistere dal commentare e diffondere le notizie che riguardano i risultati strabilianti conseguiti dalla ricerca, soprattutto quella applicativa.

Tuttavia, é anche facile cadere preda di **false suggestioni**, spesso immotivate o eccessivamente altisonanti, al punto da apparire inverosimili e irrealistiche: troppo spesso i media, ma anche gli "addetti ai lavori" (che al contrario dovrebbero manifestare un maggiore senso critico), cedono alla tentazione della propaganda.

Le motivazioni della diffusione di annunci esagerati in merito alle possibilità dell'Intelligenza Artificiale nei diversi settori della vita quotidiana (a cominciare dalle prospettive di lavoro) sono spesso ricondubili a un *business model* distorto, che **allinea gli incentivi** (anche economici e finanziari) dei **produttori di software** con quelli dei **media**: entrambi infatti hanno da guadagnare dal clamore mediatico che caratterizza non solo l'Intelligenza Artificiale, ma l'innovazione tecnologica in generale.

Pertanto, non è affatto raro oggigiorno incappare in proclami altisonanti che annunciano l'inevitabile (sopra)avvento dell'Intelligenza Artificiale, pronta a soppiantare il genere umano in ogni campo, fino al punto di spodestarlo dai residui dominii che gli sono ancora propri, quali quello della creatività e della ricerca scientifica.

Ma quanto c’é di vero, o al limite di verosimile, in tali proclami altisonanti?

In realtà poco, ma il problema é che é difficile sconfessare la **retorica dei tecno-sciovinisti** con argomenti intuitivi e convincenti, tanto fitto é l’alone di mistero che circonda tali tecnologie divenute ormai “esoteriche”, al punto che il rischio di cedere all’incantesimo della **“stregoneria digitale”** non riguarda più soltanto il cittadino comune, ma anche l’esperto del settore (o presunto tale).

PARTE PRIMA. LE ASCENDENZE CULTURALI DELLA STREGONERIA DIGITALE

LA RETORICA DELL'INEVITABILE

Se c'è un elemento che caratterizza la narrazione dell'innovazione digitale, è l'ampio uso di artifici retorici volti ad infondere nel pubblico un senso di **inevitabilità** del progresso tecnologico.

Come vedremo, tale retorica è funzionale agli interessi economici delle aziende che forniscono i prodotti e servizi tecnologici, e si allinea con gli interessi dei *media*, volti a catturare l'attenzione degli utenti.

Ma prima di andare a fondo nella nostra analisi, è utile ripercorrere brevemente l'iter storico e culturale che ha dato origine a tale narrazione, che come vedremo, trova le proprie radici insospettabili in alcune **antiche dottrine** filosofiche e religiose, riportate in auge dalla odierna **ideologia tecnocratica**.

Il primo concetto che occorre introdurre è quello di **escatologia**, che è alla base della presunta "inevitabilità" del progresso tecnologico.

L’escatologia e il destino ultimo dell’Umanità

Per comprendere le ragioni che stanno dietro alla retorica dell’inevitabilità del progresso tecnologico, occorre partire dalla “fine della Storia”, ovvero dalla dottrina che tradizionalmente va sotto il nome di **escatologia**, e che rappresenta appunto l’interpretazione che ogni tradizione filosofica e religiosa ha inteso dare al destino del genere umano.

Il termine *escatologia*, come noto, è diffuso in ambito teologico e filosofico, e riguarda gli studi dottrinari volti a rivelare i **destini ultimi** dell’Umanità, nell’intento di contribuire a chiarire il **senso esistenziale** dell’uomo.

È chiaro quindi come tale tipologia di indagine possa influire in maniera determinante sulle scelte di **condotta di vita** degli individui che si riconoscono in una determinata visione escatologica dell’esistenza.

Malgrado le analisi escatologiche siano solitamente riconducibili alle diverse concezioni religiose, esse tuttavia sono comuni anche alle dottrine filosofiche che sostengono di poter individuare un **senso nella Storia** in generale.

Caso tipico, sono le concezioni della storia avanzate dall’Idealismo tedesco ottocentesco, che vedono nella Storia uno svolgersi dialettico tra “forze” antagoniste (*tesi* e *antitesi*), che trovano la loro *sintesi* nella concreta realizzazione del *divenire* storico.

Per tali concezioni escatologiche laiche, il divenire storico è **determinato**, e come tale è destinato a realizzarsi

inevitabilmente sulla base della *"Ragione"* ultima che muove tali forze antagoniste.

Nel caso della filosofia della Storia di stampo hegeliano, la Ragione si realizza mediante l'affermazione dello **Spirito Assoluto**; nel caso del marxismo (altra dottrina della storia debitrice della concezione hegeliana), le forze del proletariato determineranno il superamento del capitalismo, realizzando l'inevitabile avvento del Socialismo Reale.

La cifra comune che caratterizza tali concezioni, sia religiose che "laiche", è rappresentata dalla **inevitabilità** dello sviluppo storico, determinato dalle forze ineluttabili che stanno dietro al "destino" rivelato dalla visione escatologica stessa.

La Tecnologia come Salvezza e Destino

Ponendosi dal punto di vista della "fine dei tempi", l'escatologia intende dare una risposta alla domanda sullo **scopo** e il **fine** dell'esistenza umana.

Essendo determinata dalle aspettative ultime riguardo al destino e agli scopi cui l'esistenza umana devono ispirarsi, le diverse interpretazioni escatologiche influiscono e **condizionano** in maniera determinante la vita e le **scelte** degli individui che si riconoscono in tali interpretazioni, e che in esse ripongono le proprie aspirazioni di **"salvezza"** e redenzione.

Di conseguenza, l'aspettativa di una vita oltramondana può indurre il credente a rinviare le proprie aspirazioni ideali (come quella di giustizia ecc.) alla **dimensione ultraterrena**.

Al contrario, una escatologia millenaristica come quella marxista, pone la dimensione salvifica all'interno della temporalità del mondo materiale, senza rinviarla ad una vita oltremondana, e tale visione informa di conseguenza le scelte e le aspirazioni di coloro i quali abbracciano tale dottrina.

Allo stesso modo, le visioni escatologiche che vedono nel **progresso tecnologico** la loro fonte di ispirazione, influiscono e condizionano in maniera determinante le scelte degli individui, sulla base delle **aspettative** (più o meno realistiche) e le concezioni del mondo, oltre che del futuro, che esse contribuiscono a diffondere.

Pertanto, se si parte dal presupposto (come fanno i visionari tecnologici) che l'**uomo** è sostanzialmente un **essere "difettoso"**

(*flawed*), e in quanto tale è bisognoso di essere **redento e mondato** dai propri *"vizi naturali"*, e che **l'unica salvezza** per l'uomo è rappresentata dalla **tecnologia**, appare chiaro che qualsiasi tentativo di contrastare (o soltanto di rallentare) il progresso tecnologico "salvifico", viene visto come un sacrilegio e un'empietà (condannati peraltro al fallimento, data l'ineluttabilità delle ragioni che governano l'inevitabile realizzazione del Progresso).

Una delle principali interpretazioni escatologiche basate sul progresso tecnologico è rappresentata dal **transumanesimo**, le cui ascendenze affondano nell'antico culto dello **gnosticismo**, come vedremo tra breve.

In realtà, non vi è proprio **nulla di "inevitabile"** nel progresso tecnologico, ma poichè (come amano spesso ripetere gli stessi "innovatori visionari") il modo migliore per prevedere il futuro è progettarlo, appare evidente come la **retorica dell'inevitabile** sia funzionale per giustificare l'adozione di determinate scelte a favore di tecnologie (pre)definite, favorendo altresì i loro proponenti.

Se a questo si aggiunge anche un'aura di "sacralità destinale", la narrazione ne risulta ancora più convincente, diffondendo così quel **"timore reverenziale"** nei confronti della tecnologia, necessario affinchè i cittadini si dimostrino disposti ad accettare (per non dire subìre passivamente) le scelte auspicate dai tecnocrati, ponendo in cattiva luce chi osa criticare e contrapporsi ad esse, qualificandolo come "retrogrado" e "irrazionale", in virtù

appunto della pretesa inevitabilità delle sorti “magnifiche e progressive” che il futuro ci riserva...

L’inevitabilità come legittimazione della Tecnocrazia

In realtà, la asserita “inevitabilità” del progresso tecnologico serve anche ad un altro scopo: ad attribuire ai tecnocrati quella **legittimità** che a loro manca, al fine di poter imporre al popolo le scelte reputate più opportune per il loro futuro.

A differenza dei *rappresentanti del popolo* eletti democraticamente, i tecnocrati trovano la loro legittimazione direttamente nella loro *expertise*: se il futuro è governato dalla **inevitabilità** del progresso tecnologico, **solo gli esperti** sono per definizione in grado di interpretare l’evoluzione del futuro ineluttabile che ci attende, e di conseguenza nella posizione di prendere le **decisioni** adeguate per assecondare “organicamente” la sua realizzazione.

Come vedremo, i tecnocrati perseguono l’antico proposito di sostituire ai rappresentanti elettivi del popolo gli **esperti** di settore, reputati più adeguati a prendere le decisioni sul futuro rispetto alla classe dirigente politica, considerata al contrario “incompetente” dal punto di vista tecnico, e come tale inadatta a gestire l’evoluzione tecnologica.

Quella della **tecnocrazia** non è altro quindi che la riedizione moderna dell’aspirazione platonica (consegnata dal filosofo greco ai posteri nel famoso libro la “Repubblica”) di insediare i **filosofi Re** al potere, in virtù della conoscenza “vera” (*episteme*), appannaggio esclusivo di tali sapienti illuminati, contrapposta all’opinione (*doxa*) che caratterizza invece il popolo.

Allo stesso modo, la riproposta di questa suggestione platonica non è altro che il tentativo di scardinare le istituzioni democratiche, con l'intento di consegnare il potere di decidere il futuro dei cittadini nelle mani degli "esperti".

IL MITO ESPONENZIALE

A suffragare la retorica dell'inevitabilità del progresso tecnologico, vi è un altro mito fondativo: quello della crescita esponenziale associata ai miglioramenti continui della tecnologia.

In questo senso, viene spesso indicata come "evidenza" a supporto la cosiddetta "legge di Moore" (che tutto è tranne che una legge, nè nell'accezione che ne darebbe la fisica, nè tantomeno nel senso giuridico del termine).

La legge di Moore nasce dall'osservazione empirica relativa alla **crescita della complessità** dei microprocessori (misurata dal numero di transistor all'interno dei chip), che **raddoppia ogni 18 mesi**, per quadruplicare di conseguenza ogni tre anni.

Alla luce di tale osservazione empirica, Gordon Moore, all'epoca capo del settore Ricerca e Sviluppo della Fairchild Semiconductor, nel 1965 ipotizzò che il numero di transistor nei microprocessori sarebbe raddoppiato ogni 12 mesi circa.

La previsione di Moore (che tre anni dopo fondò la Intel) si rivelò empiricamente corretta, e negli anni a venire mantenne la sua sostanziale validità osservazionale, contribuendo così a corroborare le aspirazioni della inevitabilità del progresso tecnologico.

Dal punto di vista formale, la "legge di Moore" rappresenta una **estrapolazione** statistica, che dalla analisi dei dati storici *noti*, estrapola appunto una "regolarità" che si assume essere valida anche per i dati futuri *ignoti*.

L’estrapolazione è un processo matematico-statistico simile a quello di interpolazione, con la differenza appunto che mentre nel caso dell’interpolazione si tenta di individuare una tendenza all’interno di un insieme di dati *noti*, nel caso dell’estrapolazione si cerca di estendere tale tendenza anche ai dati futuri, che sono per definizione *ignoti*.

Di conseguenza, le previsioni fondate sul processo di estrapolazione conservano un **elevato grado di incertezza** (come tale incompatibile con il concetto formale di “legge” rigorosa), a prescindere dal numero di conferme future che tali previsioni possono ricevere (come dovrebbero sapere bene gli investitori nei mercati azionari, per i quali vale sempre il monito che i guadagni passati non costituiscono garanzia dei guadagni futuri...)

Ma al di là della correttezza formale della cosiddetta “legge di Moore” e della relativa attendibilità delle sue previsioni future, quello che denota maggiormente il carattere *“magico”* rispetto alla pretesa inevitabilità del progresso tecnologico, è il **salto logico** compiuto dai tecnocrati nel giustificare tale inevitabilità.

Vediamo di chiarire meglio i termini della questione.

Quando la Quantità si traduce in Qualità, il corso del futuro è segnato

> “Il mutamento nella Quantità implica un mutamento nella Qualità”
> K. Marx, *“Il Capitale”*

Implicita nell’osservazione della **crescita esponenziale** della **complessità** dei microprocessori, vi è l’assunzione che la capacità computazionale che da tale crescita deriva, possa determinare l’emersione di **fenomeni “singolari”** (quali ad esempio l’emersione della *mente* e della *coscienza*) al superamento di una non meglio precisata *soglia critica*.

Senza anticipare quanto diremo più avanti in merito alla **Singolarità** attesa dai visionari alla Kurzweil, quello che ci preme sottolineare in questa sede è come tali concetti non siano altro che la versione riveduta e corretta di idee passate.

Nello specifico, intendiamo riferirci alla teoria hegeliana dello storicismo dialettico, e al materialismo storico di Marx che da esso deriva, che rappresentano concetti centrali nel determinare l’evoluzione della Storia e della Società secondo tali concezioni.

Marx fa propria la “scoperta” fatta originariamente da Hegel, che nella “Logica” aveva sostenuto che *“mutamenti puramente **quantitativi** possono risolversi a un certo punto in distinzioni **qualitative**”*.

Tale scoperta può essere definita come la **“legge del salto qualitativo”**.

In realtà, a dispetto dell’apparente cripticità, il concetto sottostante a tale affermazione è intuitivo ed è facilmente verificabile anche nell’esperienza quotidiana: basti pensare ad esempio, a come poche gocce d’acqua possano essere facilmente raccolte in un comune bicchiere, mentre miliardi di gocce si trasformino in un nubifragio, dando luogo così ad una distinzione **qualitativa** indotta da un mutamento puramente **quantitativo**.

Sempre della stessa “sostanza” si tratta (*acqua*): ma le diverse quantità coinvolte, determinano il passaggio (*salto qualitativo*) dalle gocce raccolte in un bicchiere, al nubifragio…

La novità che caratterizza la concezione dialettica è la **interpretazione metafisica** che ne viene data (prima da Hegel e poi da Marx) nel determinare il corso della Storia e della Società.

Sulla stessa falsariga dello storicismo dialettico e del materialismo storico, la **Singolarità tecnologica** è “destinata” a realizzarsi a seguito del **salto qualitativo** compiuto dalla complessità computazionale, al superamento della necessaria “soglia critica”.

Allo stesso modo, i fenonemi ritenuti “emergenti” (come la mente e la coscienza) si manifesteranno spontaneamente, al realizzarsi della Singolarità tecnologica.

Ma prima che i fenomeni “emergenti” della mente e della coscienza possano manifestarsi, occorre liberarsi del fardello obsoleto costituito dal corpo biologico.

ESCI DA QUESTO CORPO: IL CORPO MATERIALE COME GABBIA DELLO SPIRITO

Tra le ascendenze culturali che ispirano l'odierna visione "salvifica" della tecnologia, un posto di riguardo lo occupa l'antica dottrina dello **gnosticismo**.

Il termine gnosticismo deriva dalla parola greca gnósis, che può essere tradotto come "conoscenza", intesa anche nel senso di "illuminazione".

Lo gnosticismo rappresentava un movimento filosofico, religioso ed esoterico, già noto nel mondo ellenistico greco-romano, che raggiunse la sua massima diffusione tra il II e il IV secolo d.C.

L'ideale ascetico dello gnosticismo predicava l'abbandono del mondo materiale, visto come "gabbia" dello spirito.

Il mondo materiale rappresenta infatti un livello di **realtà "inferiore"** dal quale occorre liberarsi, adottando pratiche di vita che, a seconda dei culti, prevedono la povertà personale, l'astinenza sessuale, ecc.

Solo in questo modo lo **Spirito** può liberarsi dall'elemento materiale (rappresentato *in primis* dal corpo e dalla carne) che lo ingabbia, e gli **impedisce di unirsi** all'unica **vera Realtà**, quella della Divinità.

Il mondo della Divinità è al di fuori dello spazio e del tempo, e come tale **non è corruttibile**, né è soggetta ai limiti della dimensione esistenziale.

Il tipo di **conoscenza** che lo gnosticismo intendeva conseguire era di natura **esoterico e iniziatico**, volta a rifuggire dal mondo

materiale, considerato “inferiore” e impuro, per abbracciare il mondo spirituale e ricongiungersi con la divinità, rappresentata in forma impalpabile e immateriale, e come tale non soggetta ai limiti materiali spazio-temporali.

Come conseguenza di tale impostazione, il **corpo biologico** stesso è considerato come una **gabbia** che intrappola l’elemento spirituale superiore, costringendolo a permanere confinato nella realtà terrena “inferiore”.

L’eco di tale **svalutazione del corpo** materiale è rinvenibile anche nella dottrina cristiana, laddove lo stesso apostolo Paolo nella Lettera ai Romani sostiene che *“Voi non siete sotto il dominio della carne, ma dello Spirito, dato che lo Spirito di Dio abita in voi”* (Romani 8,9).

La *“carne”* rappresenta pertanto il **principio del peccato** che opera negli uomini, peccato che può essere vinto solo accogliendo la Grazia divina, attraverso la Fede, che riconduce l’uomo alla comunione salvifica con Dio.

Lo Spirito Tecnologico versus la gabbia biologica

Appare evidente l'eco delle dottrine gnostiche in gran parte delle **tecnologie "salvifiche"** e delle relative narrazioni profetiche che le accompagnano, a cominciare dal movimento del **transumanesimo**, per arrivare all'upload della mente nel cloud, passando per la "Singolarità" che a detta dei sostenitori (Ray Kurzweil in primis) darà luogo all'avvento delle Macchine Intelligenti.

Avremo modo di affrontare a tempo debito le caratteristiche di tali "narrazioni"; in questa sede ci preme sottolineare come antiche concezioni esoteriche come quelle gnostiche vengano riesumate strumentalmente per affermare la pretesa *"superiorità"* della tecnologia rispetto alla natura biologica, considerata come elemento corruttibile, dal quale occorre liberarsi.

Allo stesso tempo, la concezione della **mente computazionale** assume un ruolo centrale, prendendo il posto tradizionalmente riservato nelle dottrine gnostiche allo **Spirito "immateriale"**.

Come conseguenza, viene ribadita e rafforzata la concezione che sostiene il *dualismo "mente-corpo"*, a dispetto delle affermazioni contrarie di facciata, che solitamente caratterizzano le narrazioni ingegneristiche, ispirate al *riduzionismo* imperante nel mondo tecnologico.

LA CONCEZIONE COMPUTAZIONALE DELLA MENTE

La pervasiva diffusione dei computer e il successo conseguito nell'utilizzo delle accresciute capacità computazionali in molteplici settori della vita quotidiana, inclusa la ricerca scientifica, ha rilanciato con forza la suggestione che i meccanismi (ancora largamente ignoti) alla base della *mente* e della *coscienza*, possano essere finalmente svelati dalle macchine.

La stessa distinzione tra **"hardware"** e **"software"** ha suggerito a molti la possibilità di estendere in forma analogica la distinzione anche alla "macchina umana".

Così il *corpo* è chiamato a recitare il ruolo dell'hardware, mentre la *mente* e la coscienza non sarebbero altro che il **"software"** che viene eseguito all'interno del cervello.

Mutuando poi i concetti già introdotti di **"complessità crescente"** dei microprocessori e di **salto qualitativo**, appare immediata l'analogia tra la complessità della *struttura neurale* del cervello e le odierne **reti neurali artificiali** che proprio dalla struttura del cervello traggono ispirazione per la loro implementazione.

Se tutto questo è vero, allora sarà *solo questione di tempo* (per alcuni quel tempo è già arrivato) che le macchine non solo saranno in grado di pensare e avere coscienza di sè come gli umani, ma che la loro intelligenza sorpasserà di vari gradi quella umana, dando luogo alla **"Superintelligenza"** teorizzata da N. Bostrom, che pare essere nient'altro che la Singolarità di Kurzweil all'opera.

A supportare tale suggestione vi è la **concezione "funzionale" della mente**, che come vedremo tra breve, non è altro che il retaggio moderno delle antiche dottrine gnostiche, e al contempo la riaffermazione del *dualismo cartesiano* "con altri mezzi".

Se assomiglia a un felino, miagola e fa le fusa, allora è…uno Stregatto!

La concezione *funzionale* (o *comportamentale*) è una forma di **riduzionismo** volto a ricondurre l'essenza di un ente al suo comportamento o alla capacità di esercitare determinate funzioni.

In questo modo non solo lo Stregatto citato è rappresentato dalle sue manifestazioni esteriori, ma la stessa mente viene ridotta alle sue capacità cognitive, che si reputano essere replicabili anche da agenti software.

In sostanza, il processo di *astrazione* dalle *caratteristiche materiali* specifiche, considerate meramente *accidentali* e non essenziali ai fini delle capacità funzionali, conduce a considerare tali capacità come **autonome e indipendenti**, fino ad attribuire ad esse una loro **"identità"** specifica (nel linguaggio filosofico tale processo viene comunemente indicato con il termine *ipostatizzazione*).

Tale processo di astrazione è tipico anche del ragionamento matematico: la potenza dei numeri consiste nella capacità astratta di eseguire calcoli e operazioni (quali somma, prodotto ecc.) a prescindere dalla natura concreta degli oggetti a cui essi si applicano.

Secondo la concezione platonica della matematica (tutt'oggi in voga tra i matematici), i numeri avrebbero una loro **"realtà" indipendente** (sulla falsariga delle idee perfette esistenti nell'iperuranio platonico); allo stesso modo, la *mente* è reputata avere una propria realtà indipendente, che è riconducibile alle funzioni astratte che è in grado di manifestare.

Appare evidente l'influsso gnostico in questo processo di astrazione, con la *mente* a recitare il ruolo dello **Spirito disincarnato**.

Quello stesso Spirito disincarnato che assumerà la forma della *"res cogitans"* di Cartesio, contrapposta alla *"res extensa"* (la materialità del corpo), in cui l'individualità del pensiero è certificata dal noto *"Cogito, ergo sum"* ("Penso, dunque sono").

A dispetto dell'approccio *riduzionista* comunemente accolto in ambito scientifico, la *concezione funzionalistica* sottostante alla **"mente computazionale"** è una riaffermazione del *dualismo mente-corpo*, piuttosto che una sua dissoluzione.

In tal senso, gioca un ruolo chiave anche l'asserita **indipendenza** della mente **dal substrato materiale**, che se condotta fino all'estremo impone di considerare il substrato materiale assolutamente secondario e **non essenziale**.

L’indipendenza dal substrato materiale della mente computazionale

Molte delle suggestioni fantascientifiche avanzate dai tecnocrati risentono di questa pretesa indipendenza della mente dal substrato materiale.

Non rappresentando il corpo un elemento necessario ai fini della realizzazione della mente, essa diventa non solo *simulabile* all’interno di un computer con adeguata capacità di elaborazione (molto si fantastica sulle mirabolanti capacità degli imminenti computer *quantistici*), ma si reputa possibile persino “salvare” la mente facendone **l’upload** sul cloud!

In questo modo, si realizza una delle aspirazioni più antiche dell’uomo, insieme all’*ubiquità* e all’*eterna giovinezza*: quella di **sconfiggere la morte**, separando la coscienza (considerata alla stregua di anima “computazionale”) dalla gabbia del corpo biologico (in quanto tale “deperibile” e corruttibile) per sostituirlo con qualsiasi “supporto” materiale alternativo, rappresentato dall’hardware fungibile della macchina in silicio.

Anzi, a detta di Ray Kurzweil (tra gli entusiasti sostenitori della possibilità di effettuare concretamente l’uploading della mente), l’*emulazione* del cervello umano all’interno di un **computer** risulterebbe molto **più performante** rispetto al *“computer biologico”*, vale a dire il cervello!

In questo modo, l’eresia gnostica assume definitivamente la forma di odierna *stregoneria* digitale…

DALLA MENTE COMPUTAZIONALE ALL'INTELLIGENZA ARTIFICIALE

Se la mente stessa non è altro che un processo computazionale simulabile all'interno di un elaboratore elettronico, perchè non **ricreare computazionalmente** le capacità cognitive caratteristiche della mente, a cominciare dall'intelligenza?

È questo il passo logico successivo a quello di aver ridotto la mente a semplice processo computazionale.

In realtà, i tentativi di simulare l'intelligenza umana in forma artificiale possono essere fatti risalire fin dagli albori dell'informatica.

La stessa definizione di *"intelligenza artificiale"* si deve a J. McCarthy, che ne coniò il termine in occasione del convegno organizzato a Dartmouth nel 1956, che segna la nascita del settore di ricerca come lo intendiamo oggi.

Il programma di ricerca intendeva in una prima fase risolvere problemi di logica ben definiti, e successivamente si proponeva di emulare il comportamento umano nella soluzione di problemi di carattere generale, dando vita al filone di ricerca noto come *"Artificial General Intelligence"* (AGI).

La realizzazione di tali progetti di ricerca si è protratta fino ai giorni nostri, e attualmente prende spunto dall'emulazione di quello che si ritiene essere il funzionamento del cervello umano, replicandone artificialmente la struttura neurale, dando luogo alle attuali *"Artificial Neural Networks"* (ANN).

Tale approccio è supportato dal successo che hanno recentemente conseguito le reti neurali artificiali, in modo particolare nell'apprendimento automatizzato realizzato nella forma del **Deep Learning** (*Apprendimento Profondo*), che da più parti è ritenuto essere l'approccio più promettente per il conseguimento concreto della AGI.

I progressi conseguiti dalle reti neurali artificiali sono indiscutibili, questo anche grazie alla odierna disponibilità delle necessarie *architetture di calcolo*, che hanno permesso di implementare algoritmi sviluppati in forma teorica già nei decenni passati.

Il punto problematico che dà origine alla *"Artificial Idiocy"*, è la narrazione esagerata che di tali progressi se ne fa, arrivando agli estremi di considerare non solo come già acquisita la *Artificial General Intelligence* (cosa ben lungi dall'essere vera), sulla base della semplice *fede* nel **progresso ineluttabile** della tecnologia, ma si attribuisce a tali progressi una valenza *"salvifica"* di tutti i problemi e i guasti del mondo, attribuiti per definizione ai difetti degli esseri umani, dovuti alla loro "limitata" intelligenza (come se l'Intelligenza Artificiale non fosse essa stessa un prodotto della attività creativa umana...)

In questo senso, si intende attribuire all'Intelligenza Artificiale un **ruolo regolatore**, salvifico della stessa specie umana, individuando negli algoritmi la presenza di quel *"Nous"* che nell'antichità era considerato connotato tipico della Divinità...

L’Intelligenza Artificiale come odierno “Nous” divino

Il termine greco *νοῦς* (*Nous*) data fin dai tempi di Omero, e sta a rappresentare quella peculiare *facoltà* dell’intelletto, intesa come capacità di **comprendere gli eventi** o le intenzioni degli agenti razionali.

In Omero il termine è usato per indicare la sede della rappresentazione delle “idee chiare”, e rappresenta la capacità dell’intelletto di comprendere le “vere” **intenzioni recondite**, nonostante le “apparenze” dei comportamenti esteriori.

In tal senso, si ripropone il tema filosofico che attiene alla capacità dell’intelletto di individuare la realtà *“nascosta”* (cui si attribuisce il carattere di **“verità”**) dietro l’apparenza sensibile (caratterizzata al contrario dall’essere inaffidabile).

Anche i filosofi greci hanno conosciuto il termine sotto diverse angolazioni.

Con Anassagora il termine *Nous* assume la sua valenza metafisica più propria e completa, essendo concepito come **“Intelligenza divina”** che *organizza* il mondo.

Tale “intelligenza divina” è considerata come **potenza ordinatrice**, che dal *caos* primigenio dà origine al mondo.

Platone associerà tale intelligenza ordinatrice all’attività provvidenziale del Demiurgo, che interviene come **causa “razionale”** a plasmare la materia corruttibile ad immagine delle *idee* eterne e incorruttibili, dando origine così al *Cosmo*.

Nelle intenzioni degli odierni tecnocrati, la funzione ordinatrice che in virtù della propria "razionalità" crea ordine nel caos delle vicende umane, è da attribuirsi all'Intelligenza Artificiale, che prende così il posto del Demiurgo platonico e dell'Intelligenza "divina" di Anassagora.

In altri termini, l'**Intelligenza Artificiale** come odierno *"Nous"*, assume il ruolo di **criterio ordinatore** caratterizzato da quelle stesse prerogative ideali di razionalità, un tempo associate alle entità divine, oggi sostituite dal non meno *"sacro" algoritmo*...

IL PENSIERO MAGICO ALLA BASE DELLA STREGONERIA DIGITALE

Da quanto abbiamo detto finora, appare chiaro come la narrazione che caratterizza la tecnologia digitale si ispiri a idee tutt'altro che originali, che fanno leva anche su credenze ancestrali, che da sempre accompagnano il percorso dell'esistenza umana.

Tra queste credenze, un ruolo centrale assume il **"pensiero magico"**.

Vediamo quali sono le caratteristiche di tale forma primitiva di pensiero, e in che modo vengano riproposte in ambito *tecnologico*, per creare una narrazione più convincente facendo leva su argomenti ancestrali.

Ragionare per associazioni

Occorre premettere infatti che le forme di ragionamento umane possono assumere diverse tipologie: accanto al ragionamento *logico-deduttivo*, abbiamo infatti quello *induttivo* e *abduttivo*.

Accanto a queste tipologie di ragionamento basate sulle diverse modalità che le *relazioni causali* implicate possono assumere, abbiamo forme di ragionamento che esulano completamente da tali relazioni, e si basano invece su procedimenti **analogici**.

In altri termini, le relazioni istituite da tali argomentazioni si basano su elementi quali la **somiglianza**, la "simpatia" che li lega, ovvero la *contiguità* che caratterizza tali elementi, considerati come facenti parte di un tutto.

Tale forma di ragionamento non è solo tipico dell'uomo primitivo, ma è alla base appunto del *pensiero magico*.

Già lo stesso Frazer, nel suo famoso studio "Il ramo d'oro", sottolineò come il pensiero magico si caratterizzi per una *erronea individuazione* delle **cause**, individuando le relazioni significative tra oggetti ed eventi sulla base di **associazioni** frutto esclusivo della mente umana.

Tali forme di associazioni sarebbero istituite:

- per **somiglianza**, secondo il principio per il quale "il simile agisce sul simile" (principio che è anche alla base dell'*omeopatia*);
- per **contiguità**, sulla base del principio per cui se due elementi sono rimasti in contatto tra loro per lungo tempo, la

loro interazione si mantiene anche a distanza di tempo e luogo.

In tempi più recenti, Freud stesso accostò il pensiero magico dell'uomo primitivo a quello del bambino, sottolineando in questo modo un'altra caratteristica tipica del pensiero magico, già individuata da Frazer, che si sostanzia nella asserita **onnipotenza del pensiero**, secondo cui la realtà sarebbe influenzabile dai desideri e dai pensieri umani.

Freud estese tale caratteristica anche agli adulti affetti da **nevrosi**, i quali sarebbero inclini a dare rilevanza solo ai pensieri che implicano una intensa emotività, prescindendo dalla loro realtà oggettiva.

Magismo come forma arcana di pensiero

Malgrado studi recenti abbiano posto in dubbio che l'onnipotenza del pensiero fosse una concezione della realtà tipica dell'uomo primitivo, ancora J. Piaget nella sua "Psicologia dello sviluppo" sosteneva come il *magismo* rappresentasse una forma di pensiero arcaica, che si manifesterebbe nel bambino nella fascia di età compresa dai *due ai cinque* anni, caratterizzata dall'egocentrismo.

Solo in seguito, nelle fasi successive che conducono verso l'età adulta, il bambino si emanciperebbe da tale visione *"animistica"* della realtà, per abbracciare forme interpretative del reale basate sul ragionamento **ipotetico-deduttivo**.

Secondo Edward B. Tylor, antropologo britannico di fine '800 autore di "Primitive culture", la stessa magìa sarebbe derivata dalla concezione *animistica* della natura, tipica dei popoli primitivi, secondo cui ogni cosa avrebbe un'anima.

Tra le cose "animate" è possibile distinguere gli dèi e gli esseri *benefici* (animali inclusi), dagli esseri *malefici* (demoni e oggetti): la magìa non sarebbe altro che il modo per ingraziarsi i primi ed esorcizzare i secondi, mediante l'esecuzione di rituali basati sul potere evocativo della *parola*, in quanto espressione della già ricodata **onnipotenza** del pensiero.

Tylor fu tra i primi a parlare di magìa "simpatica", ad indicare gli effetti indotti dai **riti magici**, sulla base dei già ricordati principi *omeopatici* e di *contiguità*.

Dal magismo alla Stregoneria digitale, il passo è breve…

> “Qualsiasi tecnologia sufficientemente avanzata è indistinguibile dalla magia.”
> Arthur C. Clarke

Da quanto abbiamo detto, appare piuttosto evidente l’eco delle forme di pensiero arcaiche nella odierna narrazione che caratterizza le portentose innovazioni digitali.

In modo particolare, quelle innovazioni tecnologiche che sono legate alla cosiddetta *“intelligenza artificiale”*, alle quali è più intuitivo e addirittura spontaneo attribuire un’anima: basti pensare ai tanti *gadget* che circondano la nostra vita quotidiana, che sembrano interagire con noi con la stessa spontaneità di esseri animati…

Non solo gli *assistenti vocali*, ma sempre più i *chatbot* con cui ci relazionamo, sembrano ormai avere assunto nell’immaginario comune un ruolo finora appannaggio solo degli animali di compagnia (cui è tuttora riservata l’esclusiva dei sentimenti affettivi, ma non sapremmo dire per quanto ancora…)

Ovviamente tale rappresentazione è funzionale ai modelli di business dei produttori, che su tale suggestione fondano i propri interessi, sfruttando l’alone di mistero (di magìa, appunto) che caratterizza tali artifatti.

All’alone di mistero è anche abbinato quello della presunta **“oggettività”** degli algoritmi che sono alla base di tali tecnologie, oggettività che è direttamente ricondotta e associata alla matematica.

Matematica che tuttavia assume sempre più le sembianze di *numerologia*, man mano che l’uso delle tecnologie risponde a quello tipico del “culto del cargo”, di cui ci occuperemo tra breve.

CARGOCULTISMO TECNOLOGICO E TRIBALISMO DIGITALE

Il “culto del cargo” è un fenomeno originariamente manifestatosi in alcune tribù della Melanesia, della Nuova Guinea e della Micronesia, in occorrenza dei primi sbarchi degli esploratori occidentali, caratterizzati dall’apparizione (inedita per le tribù locali) di grandi navi e aerei da trasporto (i *cargo*, appunto) di merci e beni destinati alle popolazioni indigene, che ha assunto la maggiore diffusione a seguito del secondo conflitto mondiale, a causa dei numerosi traffici di navi giapponesi e americane nel Pacifico.

Tali osservazioni hanno indotto nelle tribù indigene a credere che i trasporti fossero opera di una divinità benevola, e quando a seguito della fine della seconda guerra mondiale furono chiuse le basi militari dell’Oceano Pacifico, cessando di conseguenza i rifornimenti di merci, i nativi credettero di potersi ingraziare la dività mediante dei riti propiziatori, affinchè i traffici di beni ricominciassero.

Il culto del cargo si sostanzia appunto in riti e pratiche religiose volte a riprodurre in forma grossolana le navi, gli aerei, le piste di atterraggio, nonchè scimmiottare i comportamenti osservati nei relativi equipaggi, nella convinzione che tutto ciò possa realizzare le condizioni necessarie ad evocare e riportare ad esistenza i fenomeni precedenti (traffici e rifornimenti di merci).

Gli indigeni fabbricavano addirittura radio e cuffie di legno per emulare i comportamenti osservati in precedenza dal personale

addetto ai trasporti, nella convinzione che costoro fossero in contatto con i propri antenati, gli unici in grado (secondo le convinzioni degli aderenti al culto) di produrre una tale ricchezza di beni.

Anche il **“culto del cargo”** si caratterizza per una forma di *magìa imitativa* basata sui principi di *associazione* simpatetica e di *contiguità* tra oggetti ed eventi, come abbiamo visto nei paragrafi precedenti.

Il termine “cargocultismo” è stato ripreso in forma ironica da Richard Feynman, il quale in un famoso discorso tenuto al California Institute of Technology in occorrenza dell’apertura dell’anno accademico 1974-75, lo utilizzò come termine di confronto tra le pseudoscienze e il metodo scientifico autentico.

Il potere “rivelatore” della Tecnologia

> “La tecnica è una forma di rivelazione.”
> Martin Heidegger

Nel mondo digitale, il cargocultismo designa la pratica di utilizzare codice di programmazione o tools in maniera pedissequa, senza che se ne sia compreso appieno il funzionamento e l’utilizzo appropriato, sperando (allo stesso modo degli indigeni inconsapevoli) di poter riprodurre gli effetti auspicati, osservati in altri contesti in cui i tools e i codici sono stati impiegati con successo.

Tale fenomeno fa emergere due elementi meritevoli di analisi:

- la **complessità** degli strumenti e delle tecnologie digitali ha raggiunto livelli tali che ormai nemmeno chi è chiamato a gestirli è in grado di comprendere appieno il loro funzionamento e le implicazioni *sistemiche* che questi strumenti comportano;
- lo stesso **“timore reverenziale”** manifestato dai non addetti ai lavori nei confronti delle tecnologie digitali di difficile comprensione, sembra pervadere anche chi dovrebbe invece essere maggiormente consapevole dei pregi e dei difetti di tali tecnologie.

L’effetto “alone” (*halo effect*) determinato dalla combinazione di complessità e “oggettività” matematica che caratterizza le procedure algoritmiche, induce la maggioranza degli utilizzatori a

riporre la propria **fiducia** (meglio sarebbe parlare di “fede”) in maniera **incondizionata** e **acritica** in tali strumenti, anche in relazione ad un rapporto di **sudditanza psicologica** determinato non soltanto dall’utilità pratica degli strumenti, ma soprattutto dallo stupore di fronte al potere *“mostruoso”* che essi **rivelano**, in relazione alla loro natura *“sublime”*, come già Kant ebbe modo di osservare a suo tempo 1.

Come correttamente fa notare l’Autrice del saggio citato: *“La radice del sostantivo ‘mostro’ (monstrum), derivato da ‘monere’, cioè ammonire, avvisare, tiene insieme il significato di avviso, annuncio di qualcosa che è fuori dall’ordinario, anzi contro l’ordine naturale delle cose, con quello di far vedere, esporre, condividendo la radice con il verbo ‘mostrare’.*

Allo stesso modo, anche l’etimologia del sostantivo *prodigio* (prodigium) sottolinea il fatto che qualcosa viene *mostrato*; in altri termini, qualcosa di inaudito ed estraneo al consueto, si **rivela** manifestando nella sua interezza il proprio **potere sovversivo** dell’ordine naturale delle cose, inducendo di conseguenza nello spettatore quel timore reverenziale, misto a stupore, di cui abbiamo parlato.

Associato al potere *“rivelatore”* della tecnologia, vi è quello *“liberatorio”*, che è stato anticipato dalla controcultura degli anni ’60 che ha messo le radici proprio in quella terra in cui vedranno la luce le maggiori innovazioni digitali degli ultimi decenni…

1. cfr. “Dal sublime al mostruoso. Due letture kantiane” di Daniela Angelucci, disponibile al link: https://journals.mimesisedizioni.it/index.php/studi-di-estetica/article/view/919↩

TI SOGNO CALIFORNIA…

La Silicon Valley e il suo ecosistema di aziende innovative basate sull'impiego delle tecnologie digitali, non sorge di punto in bianco dal nulla, ma affonda le proprie *radici ideologiche e culturali* in quella controcultura che negli anni '60 prese piede in California, e di lì si diffuse in tutta l'America.

Epicentro di tale "rivoluzione" culturale, manco a farlo apposta, è proprio l'Università californiana di Berkeley, che rappresenterà il punto di riferimento intellettuale dei **movimenti libertari** dell'epoca.

Vediamo di ripercorrerne le vicende salienti, a partire da quelle di uno dei primi visionari epigoni della Silicon Valley: Stewart Brand.

Brand è stato un personaggio poliedrico, che con le sue attività e iniziative ha ispirato molte delle innovazioni tecnologiche a venire (a cominciare dello stesso *World Wide Web*).

Dotato di spiccate capacità organizzative, è stato capace di incanalare le aspirazioni ideali e spirituali della sua generazione, facendole confluire nelle potenzialità tecnologiche della nascente industria dei personal computer, contribuendo a rivelare il loro **potenziale libertario** e di emancipazione dell'individuo, in contrapposizione al controllo posto in atto dal governo centrale americano, nei confronti del quale andava crescendo l'insofferenza ormai manifesta dei movimenti giovanili degli anni '60.

Tale *libertarismo* in chiave di enancipazione ha dato corpo a quel sogno di trasformazione del mondo, che vedeva nella tecnologia la "cura" ai guasti della società, incentivando un modello di

collaborazione globale, che prendeva spunto dalle esperienze comunitarie delle “comuni” degli hippie della Summer of Love.

La visione ingenua della funzione “salvifica” attribuita alla tecnologia verrà interiorizzata dalla nascente Silicon Valley, e ispirerà gran parte dell’ideologia (e della retorica) che la caratterizzerà negli anni a venire.

Antesignana del Web sarà la pubblicazione del “Whole Earth Catalog” a opera di Brand, che nei suoi quattro anni di esistenza arriverà a vendere 2,5 milioni di copie, vincendo persino un National Book Award.

All’interno del catalogo erano elencati una varietà di attrezzi, libri e in generale tutti gli strumenti migliori disponibili, per consentire all’utente di poterne disporne in autonomia per i propri scopi (in questo senso, il catalogo anticiperà le funzioni tipiche del futuro world wide web, come già ricordato).

I beni disponibili nel catalogo erano corredati non solo da prezzi e fornitori, ma anche da analisi e spiegazioni sul loro utilizzo; ma ancora più importanti dei beni stessi e delle informazioni su di essi, erano le argomentazioni a supporto della “filosofia” di fondo che ispirava il catalogo: ovvero permettere all’individuo di curare in autonomia la propria educazione (anticipando quelli che saranno i futuri MOOC) modellando il proprio ambiente circostante e condividendo la propria esperienza con chiunque altro fosse interessato.

Un vero concentrato di libertarismo individualista, con tanto di “self-empowerment” tecnologico!

In questo senso, la tecnologia (in modo particolare i computer) ricoprirà un ruolo centrale, e il Whole Earth Catalog stesso rappresenterà la trasposizione dei valori della controcultura giovanile nella nascente industria dei personal computer.

La Tecnologia: da strumento di oppressione a mezzo di emancipazione

Occorre dire che all'epoca i computer erano per lo più considerati strumenti di controllo funzionali alle politiche oppressive del Governo, in quanto espressione del potere burocratico e di controllo sociale che essi consentivano.

All'epoca inoltre i computer erano identificati con il loro produttore principale, la IBM, che in virtù del potere monopolistico detenuto sul mercato (alla fine degli anni '50, l'IBM controllava il 70% del mercato), insieme al sostegno del Pentagono e di altri rami dello stato, contribuiva a dare un'immagine di **centralizzazione del controllo** politico sui cittadini (in maniera non dissimile dal sistema politico antagonista a quello americano, vale a dire il regime sovietico).

Brand condivideva tale visione generale sul ruolo e la funzione dei computer, ma era persuaso del fatto che se la tecnologia era all'origine dei mali del mondo, la stessa tecnologia poteva rappresentare la **"soluzione" ai difetti** che essa stessa contribuiva a creare (in questo modo offrendo argomenti alla retorica ideologica dei futuri Big tecnologici dell'industria del *personal computing* a sostegno del "people empowerment" per via tecnologica).

Occorreva quindi sottrarre la tecnologia dalle mani del governo e dei monopolisti, per renderla disponibile agli individui (in linea con quello spirito di condivisione che ispirava le comuni degli hippie), abilitando in questo modo una singolare fusione tra una

concezione di individualismo radicale e libertarismo, che costituirà la cifra specifica della futura Silicon Valley.

In questo senso i computer, da strumento di controllo e oppressione, invenzione delle grandi istituzioni al servizio del "sistema", divenivano **strumenti di liberazione personale** e di connessione comunitaria.

Con le sue idee e iniziative, Brand ha contribuito alla creazione di un'immagine *epica e gloriosa* dell'informatica, perpetuando i valori e gli ideali del *comunitarismo* della controcultura hippie: ciò che le comuni non erano riuscite a realizzare, i computer lo avrebbero completato.

Ma la filosofia del Whole Earth Catalog andava anche oltre, ispirando una concezione "ecologica" (in cui tutto era legato a tutto il resto, in linea con la futura ideologia della rete come "organismo vivente"), oltre che di "villaggio globale".

È qui che la storia di Brand si incrocia con un altro eminente visionario dell'epoca: il sociologo canadese Marshall McLuhan.

La rete come “villaggio globale”

> “Oggi, dopo più di un secolo di tecnologia elettrica, abbiamo esteso il nostro stesso sistema nervoso centrale in un abbraccio globale, abolendo sia lo spazio che il tempo per quanto riguarda il nostro pianeta”.
> Marshall McLuhan

Nelle sue opere, McLuhan aveva predetto che le nuove tecnologie avrebbero potuto collegare il mondo in una rete, contribuendo a sanare la frattura che connotava una generazione frammentata dalla minaccia di una guerra nucleare, da un lato, e dalla alienazione sociale, dall’altro.

McLuhan attribuiva addirittura all’invenzione della stampa a caratteri mobili, introdotta da Gutenberg, la colpa della stessa frammentazione sociale, che si estrinsecava nell’atto stesso della lettura, reputato “egoistico”.

Secondo McLuhan infatti, l’alfabeto rappresenterebbe una tecnologia di frammentazione sociale, che darebbe luogo ad un “deserto” di dati classificati, a differenza della cultura orale, che presupponeva e incentivava le interazioni sociali *faccia a faccia*.

La tecnologia però forniva una possibile soluzione: i computer avrebbero riportato in auge i fasti della cultura passata, stavolta su scala planetaria, mettendo in comunicazione tra di loro in tempo reale gli utenti della rete, che avrebbero così potuto prendere

parte a quel *“villaggio globale”* alternativo alla realtà frammentata dell’individuo, confinato all’interno dei limiti spazio-temporali della propria geografia locale.

Il superamento della frammentazione darà origine anche a fenomeni olistici inediti: uno tra questi è l’emersione della cosiddetta **“intelligenza collettiva”**.

L'INTELLIGENZA COLLETTIVA DELLA RETE

Gli ideali comunitari ispirati dalle esperienze delle "comuni" degli hippie trovano la loro eco in un altro fenomeno "emergente" della rete: la cosiddetta *Intelligenza Collettiva*.

Il concetto di Intelligenza Collettiva è stato analizzato e reso noto dallo studioso francese Pierre Lévy, che ad esso ha dedicato il testo "L'intelligenza collettiva. Per un'antropologia del cyberspazio" del 1994.

Tuttavia, come spesso capita, l'idea dell'esistenza di un'intelligenza che emerge al di sopra dei singoli individui non è nuova, e anche in questo caso, è possibile ritrovare negli scritti di Karl Marx un concetto analogo, quello di "intelletto generale", elaborato negli anni 1857-58 nei Grundrisse.

Per Marx, *l'intelletto generale* rappresentava una manifestazione *astratta* del **lavoro sociale**, che traeva sostanza dalla conoscenza impersonale diffusa e sedimentata all'interno del corpo sociale e del suo retroterra culturale, che esprimeva le capacità creative della società emergenti dalla collettività degli individui che la compongono.

La novità rappresentata dalla Intelligenza Collettiva, come analizzata da Pierre Lévy era costituita dai moderni mezzi di comunicazione che la rete internet rendeva disponibili, abilitando i singoli individui a condividere le proprie conoscenze in una specie di *intelligenza distribuita*, che contribuiva ad emancipare i singoli dai loro limiti individuali (come ad es. la memoria nozionistica).

Il motto di Lévy era condensato nelle affermazioni *“nessuno sa tutto, ognuno sa qualcosa”* - e *“la totalità del sapere risiede nell’umanità”*.

In quanto fenomeno “emergente”, l’Intelligenza Collettiva manifestava caratteristiche olistiche che non erano riducibili alle singole componenti, in virtù della nota proprietà caratteristica dei sistemi complessi, secondo la quale *“il tutto è maggiore della somma delle parti”*.

In questo senso, molte erano le analogie con altri fenomeni emergenti osservabili in natura, come ad esempio l’organizzazione sociale delle api, che presto diventerà, nell’immaginario fanta-politico dei sostenitori dell’Ingelligenza Collettiva, l’esempio paradigmatico da seguire anche nell’innovativa organizzazione della società umana resa possibile dall’avvento della rete internet.

Intelligenza Collettiva e la Saggezza della Folla

Ma il fondamento teorico che più di ogni altro sembrava dare sostanza alla Intelligenza Collettiva, era rappresentato dalla cosiddetta **“Saggezza della Folla”**, oggetto di analisi in un famoso saggio di James Surowiecki del 2005, dal titolo “The Wisdom of Crowd”.

La *saggezza della folla* traeva origine da un’osservazione empirica condotta da Francis Galton (cugino del più noto Charles Darwin), il quale in un articolo pubblicato nel 1907 nella rivista Nature, descriveva come la stima del peso dei buoi fornita dalla folla di astanti a una fiera di bestiame, si era rivelata più accurata rispetto alle stime formulate dai singoli esperti.

La mediana delle stime fornite dagli astanti era infatti più aderente al reale valore, rispetto alle stime fornite dagli esperti.

Sulla base di tale osservazione, veniva elaborata la teoria sociologica della Saggezza della Folla, secondo la quale, appunto, le valutazioni e le opinioni espresse da una comunità indistinta di indvidui senza particolari *expertise* nel dominio di competenza, risultavano essere più adeguate, nel loro insieme, rispetto a qualsiasi parere espresso dagli esperti del settore.

Tale teoria sociologica, unita alle innovative tecnologie digitali del Web 2.0 (vale a dire, le tecnologie del web che consentivano la condivisione delle conoscenze degli utenti, mediante strumenti tecnologici rappresentati da blog, wiki, fino agli odierni social media), davano corpo e sostanza alla Intelligenza Collettiva, come la conosciamo oggi.

Tra gli esempi di successo concreti della Intelligenza Collettiva in ambito digitale, i suoi sostenitori solitamente ricordano Wikipedia, la famosa “enciclopedia” condivisa realizzata dagli utenti della rete, e i software *open source* (Linux in primis).

Vedremo nel prosieguo della trattazione i limiti teorici che sono alla base della presunta Intelligenza Collettiva; in questa sede ci interessa descrivere le varie ascendenze culturali, rinviando alle sezioni successive l’analisi critica di tali ascendenze, e gli effetti negativi che esse determinano nei confronti del “senso comune” e del dibattito pubblico.

E tra le suggestioni pericolose che partendo dall’Intelligenza Collettiva si spingono ben oltre il ragionevole, non poteva certo mancare l’ipotesi della Coscienza della Rete.

LA COSCIENZA DELLA RETE

Dalla Intelligenza Collettiva alla Rete Consapevole di se stessa il passo è breve: anche in questo caso, sembra essere in azione una forma di pensiero simpatetico, che sfrutta l'apparente e suggestiva **analogia** tra le connessioni neurali del cervello umano, dalle quali si suppone emerga la "coscienza", e le intricate e complesse connessioni che caratterizzano la rete internet.

Facendo leva sulla tesi del **"salto qualitativo"** di cui abbiamo parlato in precedenza, si dà per assodato che al superamento di una non meglio precisata "soglia critica" di complessità delle connessioni all'interno della rete internet, essa possa manifestare fenomeni "emergenti" alla stessa stregua del cervello umano, a cominciare dai fenomeni olistici del pensiero e della mente, fino ad arrivare alla coscienza.

Che tale ipotesi sia suggestiva non c'è alcun dubbio, e la sua forza di convincimento poggia su tutto il corredo di argomentazioni che abbiamo analizzato finora.

Del resto, il terreno culturale per la diffusione delle tecnologie "cognitive" era stato preparato già da tempo.

L'idea stessa che la **tecnologia digitale** costituisca una **"estensione cognitiva"** del cervello umano rappresenta ormai un luogo comune accettato come vero da molti commentatori.

Non solo gli smartphone, ma anche gli stessi servizi web, a cominciare dai motori di ricerca per finire ai navigatori, costituiscono delle **"protesi" delle facoltà cognitive** naturali, in primis della memoria, di cui ormai sembra non si possa fare più a

meno (basti pensare al numero crescente di utenti che hanno difficoltà a ritrovare la strada di casa senza l'aiuto del navigatore…)

Lo stesso McLuhan è stato profetico nel descrivere la tecnologia come un'estensione del nostro cervello, ma a dare man forte alla ipotesi della possibilità che non solo le macchine, ma la rete internet stessa nel suo complesso, possa manifestare *fenomeni emergenti* come la coscienza, è stata la **concezione funzionalista** della mente, e il suo corollario, ovvero l'indipendenza delle facoltà cognitive dal substrato materiale.

Tale asserita **indipendenza dal substrato** implica che il materiale biologico non rappresenti una precondizione necessaria per replicare fenomeni emergenti come la mente, o come la vita stessa; ma che, al contrario, esso possa essere validamente sostituito dal silicio.

In tal senso, la scoperta e la progressiva decodifica del "codice della vita", vale a dire del DNA, ha dato ulteriore spinta alla possibilità che la realtà materiale stessa sia riducibile ad **informazione**.

E se la realtà nel suo complesso (sia essa *materiale*, che *immateriale*, come nel caso delle funzioni cognitive superiori), è riducibile ad informazione, ne segue che la realtà non è soltanto **rappresentabile** e **simulabile**, ma anche **riproducibile** in forma **computazionale** mediante gli opportuni *algoritmi*.

Tutto questo ci porta direttamente all'idea distopica di connettere direttamente i cervelli tra di loro, trasfondendoli all'interno della Singolarità.

LA SINGOLARITÀ E I CERVELLI CONNESSI

Il cerchio si chiude dunque dove tutto era iniziato, ovvero con la definitiva consacrazione della eresia gnostica attualizzata nella *Singolarità*, vale a dire la visione distopica propugnata da personaggi quali Ray Kurzweil che favoleggiano di una possibilità (che come al solito, ha il carattere di *inevitabilità*) che le macchine superino quel livello critico di complessità che le porterà a fare il "salto di qualità" che le renderà finalmente *coscienti*.

Tuttavia la Singolarità non si limita a questo, ma identifica come propria "logica" conseguenza la possibilità di connettere tra loro cervelli umani e macchine "coscienti", a formare un'unica **"Coscienza globale"** nella quale saranno immersi e ove confluiranno i pensieri e le esperienze dei singoli individui.

Appare evidente l'aspirazione **gnostica** sottostante all'idea della Singolarità, e alla conseguente possibilità di riunirsi finalmente all'unica Divinità concepibile dalla dimensione postumana, quella algoritmica dell'Intelligenza Artificiale (intesa come *Nous*) ormai unicamente e definitivamente appannaggio delle macchine.

Per quanto bizzarra possa apparire tale eventualità, occorre dire che tentativi concreti di "connessione" dei cervelli umani alle macchine sono già attualmente in corso, basti pensare al progetto Neuralink di Elon Musk [1] che si propone l'obiettivo di aiutare persone affette da malattie neurologiche a comunicare direttamente con un device protesico attraverso il pensiero.

A prescindere dall'intento meritorio di riabilitare le capacità perdute dai pazienti, il progetto lascia aperte non solo molte

questioni etiche, ma anche quesiti attinenti i possibili abusi in ottica di “capitalismo della sorveglianza”, per citare la felice definizione di S. Zuboff, e le ulteriori limitazioni dei diritti civili che ne possono conseguire.

Ad oggi Neuralink non “legge la mente”, ma interpreta la volontà di attivare i device connessi al cervello sulla base della interazione con le funzioni linguistiche che si attivano quando l’utente si limita semplicemente a *pensare* di parlare.

In altri termini, è ancora l’utente che deve attivare consapevolmente il device connesso.

Sarebbe tuttavia ipotizzabile che la macchina possa ricostruire i pensieri (anche indirettamente) senza il consenso o la consapevolezza dell’utente?

Porsi una domanda del genere non significa già riconoscere dignità e credibilità ad una visione distopica che dà per acquisita la possibilità in linea di principio di una tale “lettura del pensiero” da parte della macchina?

L’idea della Singolarità condensa quindi diverse visioni distopiche che vanno dal passaggio definitivo dall’umanità alla *post-umanità*, contribuendo a fornire ai post-umani quella dimensione teologica che mancava, insieme alla trascendenza di matrice tipicamente gnostica, che sfocia nella fusione delle menti in un’unica Coscienza globale.

Ma rappresenta anche l’ideale politico di approdare finalmente ad una **dimensione collettivistica** totalizzante.

Non a caso, l’ipotesi distopica della Singolarità era già stata vagheggiata in altri termini dai “cosmisti” sovietici.

Postumanesimo in salsa sovietica

Ai tempi dell'ex Unione Sovietica, era già presente una singolare mistione di idee che univa tra loro la spiritualità di matrice tipicamente gnostica, al materialismo dialettico.

Tale corrente di pensiero aveva assunto negli anni venti del '900 il nome di **"biocosmismo"**, che si diffuse solo durante il primo e l'ultimo decennio del regime sovietico.

Le aspirazioni manifestate dai biocosmisti si riassumevano nell'intento di realizzare in terra le promesse ultramondane delle religioni tradizionali.

La via per conseguire tali promesse era rappresentata, manco a dirlo, dalla tecnologia, i cui progressi avrebbero consentito la realizzazione del "paradiso terrestre" collettivista, superando non solo le sofferenze umane esistenti, ma spingendosi anche a conseguire l'immortalità e la resurrezione dei morti.

Allo stesso modo, sarebbero state abolite le differenze di genere sessuali (tesi che sembra prefigurare gli odierni movimenti di rivendicazione dei diritti LGBTQI+) e la procreazione sarebbe stata affidata alle procedure biotecniche di riproduzione.

Tutti gli esseri viventi, animali inclusi (anche in questo caso prefigurando i movimenti animalisti futuri) avrebbero contribuito alla realizzazione di una Ragione collettivizzata che non si sarebbe limitata al pianeta Terra, ma avrebbe coinvolto l'intero cosmo...

Lo stesso Trockij ebbe a dire che *"Produrre una nuova, «versione migliorata» dell'uomo: è quello il compito futuro del comunismo"*,

superando così le molte contraddizioni della natura umana dovute alla sua evoluzione naturale spontanea, non pianificata.

Non vanno sottaciute le implicazioni *teologiche* della visione *cosmista*, che aspirano alla realizzazione di una consapevolezza collettiva che superi le individualità dei singoli basate dalle barriere biologiche rappresentate in primis dai *corpi* e dalla riproduzione sessuale, che nell'ottica capitalistico-borghese rappresenta l'anelito all'immortalità tramite la sopravvivenza e la conservazione riproduttiva della specie (il *"crescete e moltiplicatevi"* delle Sacre Scritture), laddove nel comunismo realizzato tale anelito viene superato attraverso la vita comune del nuovo essere post-umano asessuato.

Tutto questo insieme di suggestioni tecnologiche si basava (e si basa tuttora) su una commistione di credenze, suggestioni, ideologie, miste a scoperte scientifiche (in alcuni casi autentiche, ma decontestualizzate), che si pretende possano vantare la stessa autorevolezza e attendibilità riconosciuta alla Scienza autentica, intesa come sapere critico, che tuttavia non pretende di assolvere a un intento "salvifico", e soprattutto non riconosce come proprio un sapere "dogmatico".

Ma è proprio da tale interpretazione salvifica e dogmatica della tecnologia e della scienza che ha origine quella che in tono enfatico, ma riteniamo decisamente appropriato, definiamo **"Stregoneria digitale"**.

---

1. cfr. https://neuralink.com/↩

PARTE SECONDA. SMASCHERARE L'ARTIFICIAL IDIOCY

INTRODUZIONE ALLA PARTE SECONDA

Dopo avere introdotto in forma sintetica le ascendenze culturali che hanno dato luogo alla narrazione salvifica che caratterizza la *Stregoneria Digitale*, in questa parte del testo analizziamo le forme in cui essa si manifesta concretamente come *Artificial Idiocy*, dovendosi intendere con questo termine la irrazionale ed eccessiva fiducia riposta dagli umani nelle supposte capacità "taumaturgiche" della intelligenza artificiale nel risolvere i mali dell'umanità.

In questo senso, la *Artificial Idiocy* consiste nella pretesa irragionevole di poter abdicare dalle responsabilità "umane, troppo umane", attribuendo alla tecnologia un ruolo che ontologicamente non le compete.

Di seguito prenderemo in considerazione alcuni *case studies* smascherando gli inganni e gli artifici retorici utilizzati non solo dagli addetti *marketing* delle aziende produttrici, ma sempre più dai *media*, che giocano ormai un ruolo tutt'altro che secondario nella diffusione del clamore ingiustificato che caratterizza la cosiddetta Intelligenza Artificiale, e la sua asserità "superiorità" rispetto all'intelligenza naturale (superiorità che viene data ormai per acquisita in virtù di un "atto di fede" dei tecnocrati).

LE SUGGESTIONI ALCHEMICHE DELL'ARTIFICIAL IDIOCY

All'alba di quello che sarà ricordato come il primo "inverno dell'Intelligenza Artificiale" (*AI Winter*), vale a dire la reazione fisiologica di frustrazione e scetticismo che fa seguito necessariamente alle aspettative eccessive, alimentate dai media e dai sostenitori più irrudicibili, molti commentatori e appartenenti al mondo accademico avevano iniziato a dubitare seriamente della scientificità del progetto di ricerca inaugurato con tanto clamore da McCarthy in occasione del famoso Summer Workshop del 1956, tenutosi a Dartmouth 1.

Tra i più critici ci sarà Hubert Dreyfus, che nel rapporto commissionatogli dalla RAND Corporation a metà anni 1960 per stilare lo stato di avanzamento della ricerca in ambito Artificial Intelligence, senza mezzi termini farà ricorso a concetti quali "alchimia", per denotare in negativo lo statuto di (non) scientificità da assegnare alla recente disciplina, intitolando appunto il suo rapporto come "Alchemy and Artificial Intelligence".

Anche a causa delle promesse eccessivamente ottimistiche (e irrealistiche) così entusiasticamente propagandate e date sostanzialmente per acquisite dai ricercatori dell'AI, era inevitabile che la credibilità del settore ne avesse a risentire (e con essa anche i finanziamenti), nel momento in cui si fosse arrivati alla resa dei conti con la realtà dei fatti.

Nell'ambito della comunità scientifica, l'Intelligenza Artificiale era ormai considerata alla stregua delle pratiche pseudo-scientifiche, e

per essa si apriva il primo “inverno”, finendo relegata per circa un decennio, che va dai primi anni ’70 agli inizi anni ’80, nel dimenticatoio dei tanti programmi di ricerca falliti…

Paragonare un disciplina all’alchimia è infatti considerato sommamente offensivo in ambito scientifico: per molti rappresentava la pietra tombale sull’intero settore di ricerca.

Ma come è stato possibile passare così rapidamente “dalle stelle alle stalle”?

Per descrivere il passaggio traumatico dagli iniziali entusiasmi alle inevitabili delusioni, possiamo analizzare la storia delle fortune e sfortune di uno dei primi modelli predittivi di intelligenza artificiale che sia stato mai implementato: il Perceptron di Rosenblatt.

Ma prima di esaminare il modello del Perceptron, dobbiamo capire come è nata l’idea di realizzare l’Intelligenza Artificiale prendendo come spunto la struttura del cervello biologico.

La logica alla base della matematica e dei neuroni

I primi anni del ’900 sono stati caratterizzati da un notevole fermento di ricerca in ambito logico-matematico: sulla scorta del programma di ricerca inaugurato da G. Frege, volto a ricondurre la matematica ad un nucleo fondamentale di assiomi e regole logiche, Bertrand Russell e Alfred North Whitehead publicheranno negli anni che vanno dal 1910 al 1913 l’opera monumentale nota come *“Principia Mathematica”*.

L’intento degli Autori era di pervenire ad una sistematizzazione dei fondamenti logici della matematica, superando anche alcune aporie (note come “paradossi”) a causa delle quali il lavoro di ricerca di Frege si era arenato.

L’idea che la matematica potesse ridursi alla logica aveva implicazioni filosofiche straordinarie: da un lato testimoniava della potenza esplicativa della logica formale, dall’altro attribuiva carattere di universalità, verità e certezza “matematica” a tutte le proposizioni e asserzioni derivanti dalla corretta applicazione delle regole della logica, sulla base di principi primi fondati sull’evidenza osservativa.

Occorre infatti ricordare che in quegli anni la comunità scientifica era alla ricerca di un **principio di “verificazione”** che desse conto della affidabilità dei risultati conseguiti dalla applicazione del metodo scientifico, in contrapposizione alle “fantasie metafisiche”, sostanzialmente sterili, che caratterizzavano invece la filosofia.

Il paradigma di riferimento della ricerca scientifica dell'epoca era ancora impiantato sul **"positivismo logico"**, che asseriva la possibilità da parte della conoscenza scientifica di conseguire il più elevato grado di affidabilità (la certezza) delle proprie scoperte, mediante l'analisi logicamente fondata condotta sui fatti "positivi", vale a dire sulle osservazioni empiriche non inquinate da pregiudizi metafisici e astrazioni aprioristiche, retaggio di antiche pratiche considerate antiscientifiche.

Al contempo, la nascente area di ricerca delle neuroscienze poneva le basi per riconsiderare l'attività del cervello biologico alla stregua di un'attività di elaborazione realizzata dai neuroni, sulla base di rigorose regole logiche.

Sarà quindi lo storico incontro tra due delle menti più originali e innovative a gettare le basi del primo modello computazionale su base neurale.

Quando il biologo incontra il logico

L'incontro in larga parte fortuito e rocambolesco, caratterizzato da alterne fortune, tra Walter Pitts, autentico genio della logica, che fin dall'età di 12 anni mostrerà il suo talento scovando nella monumentale opera di Russell e Whitehead alcuni errori (si racconta che Pitts lesse per intero i "Principia Mathematica" dopo essersi rifugiato in una biblioteca ed esservi rimasto tutta la notte per sfuggire a dei bulli che lo molestavano), e Warren McCulloch, fisiologo di professione, interessato allo studio del cervello e del sistema nervoso, darà luogo allo storico paper dal titolo 'A logical calculus of the ideas immanent in nervous activity' pubblicato nel 1943.

Per la prima volta si metterà in relazione la struttura biologica e le funzioni svolte dai neuroni con gli operatori logici e le regole di inferenza della logica di Boole (anche nota come "logica binaria", che è alla base degli odierni computer).

Esaminando la biologia dei neuroni sulla base delle conoscenze dell'epoca (secondo le quali i neuroni hanno corpi cellulari e assoni, che consentono di connettere tra loro altri neuroni, e attraverso questa connessione fornire un input da un neurone all'altro, il quale si attiva al superamento di una determinata soglia elettro-chimica), McCulloch e Pitts sono in grado di dimostrare la congruenza esistente tra i **neuroni biologici** e le regole della **logica binaria**.

L'elemento che consente di istituire il parallelo tra biologia e logica è rappresentato dal fatto di considerare lo stato del neurone

“attivato” come sinonimo della condizione logica “vero”, laddove il neurone inattivo è rappresentativo dello stato logico “falso”.

Combinando insieme diversi neuroni in stati logici determinati, è possibile ricreare dei “circuiti” in grado di rappresentare qualsiasi operazione di logica binaria.

Questo risultato teorico avrà un’importanza fondamentale nei futuri progetti di ricerca volti a ricreare artificialmente i meccanismi logici che fino ad allora erano considerati esclusivo appannaggio della mente umana.

La realizzazione pratica del primo neurone artificiale costituirà motivo di straordinario entusiamo in relazione alla possibilità concreta di creare un “cervello elettronico”.

Potevamo stupirvi con Percettroni strabilianti

Esibendo una capacità comunicativa e istrionica tipica di un piazzista navigato, il dott. Frank Rosenblatt presenterà la sua creatura, il **Perceptron**, come "l'embrione di un computer elettronico che, una volta completato in circa un anno, dovrebbe essere il primo meccanismo non vivente in grado di 'percepire', riconoscere e identificare ciò che lo circonda senza addestramento o controllo umano".

Secondo Rosenblatt, al tempo ricercatore in psicologia presso il Cornell Aeronautical Laboratory, il Perceptron "sarà il primo dispositivo elettronico della storia a pensare come un cervello umano, e come un cervello umano, imparerà dai suoi errori", e la notizia sarà enfaticamente riportata nell'articolo intitolato 'Electronic "brain" teaches itself' apparso sul New York Times del 13 luglio 1958.

Il termine Perceptron designa una classe di dispositivi in grado di individuare somiglianze e **classificare** di conseguenza gli oggetti, sulla base di un addestramento effettuato su dati di ingresso forniti da immagini fotografiche scattate da una telecamera montata su un treppiede posto all'estremità della macchina.

Come i modelli introdotti nell'articolo di McCulloch-Pitts, anche il Perceptron era una rete neurale artificiale, che intendeva replicare in forma semplificata le connessioni e le funzioni dei neuroni biologici.

A differenza dei modelli teorici di McCulloch-Pitts, che erano rimasti sulla carta, il Perceptron intendeva essere una

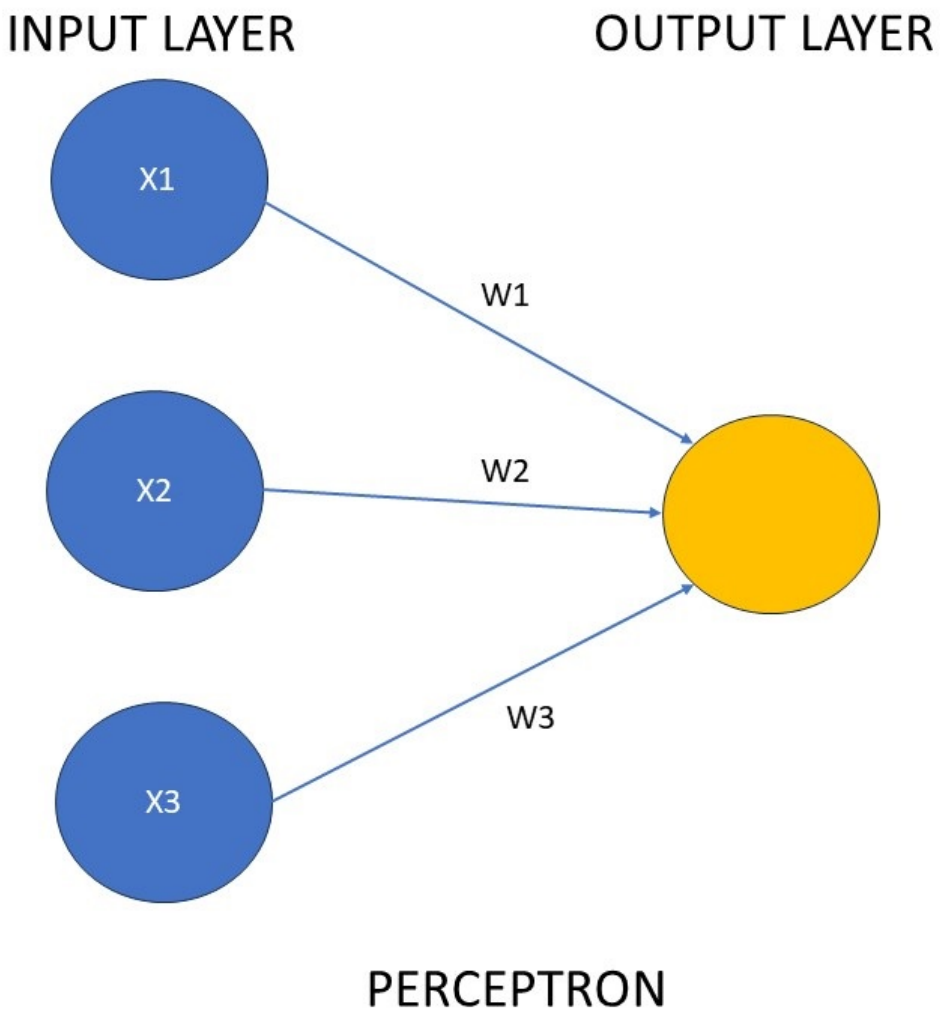
implementazione concreta e funzionante di una rete neurale.

Inoltre, il Perceptron era stato progettato per **apprendere autonomamente** a classificare i dati in ingresso, attribuendo un **peso differente** alle diverse connessioni all'interno della rete neurale (a differenza dei modelli teorici di McCulloch-Pitts, che si limitavano a stabilire la funzione logica istituita dalle connessioni all'interno della rete), sulla base degli esempi forniti nella fase di addestramento.

Il Perceptron è in grado di apprendere seguendo l'**approccio supervisionato**: in altri termini, ha bisogno di dati di addestramento previamente etichettati (classificati) da assumere come esempi di input, sulla base dei quali correggere gli eventuali errori di predizione.

Il Perceptron pertanto è in grado di apprendere autonomamente **correggendo progressivamente** gli **errori di predizione** attraverso la modifica dei **pesi attribuiti** di volta in volta alle singole connessioni all'interno della rete neurale, per adattare le predizioni ai risultati richiesti.

Nell'immagine che segue, viene mostrato il modello della rete neurale impiegata dal Perceptron, in cui ai diversi input (identificati con i valori `x1`, `x2`, ecc.) vengono associati i relativi 'pesi' (indicati come `w1`, `w2`, ecc.) idonei per ottenere un determinato output:



Perceptron di Rosenblatt

Nel modello originario implementato da Rosenblatt, il Perceptron associava ai diversi ingressi della telecamera rappresentati dai sensori, altrettanti collegamenti della rete neurale, costituiti da circuiti elettrici che si attivavano in corrispondenza dell'input ricevuto dai sensori stessi.

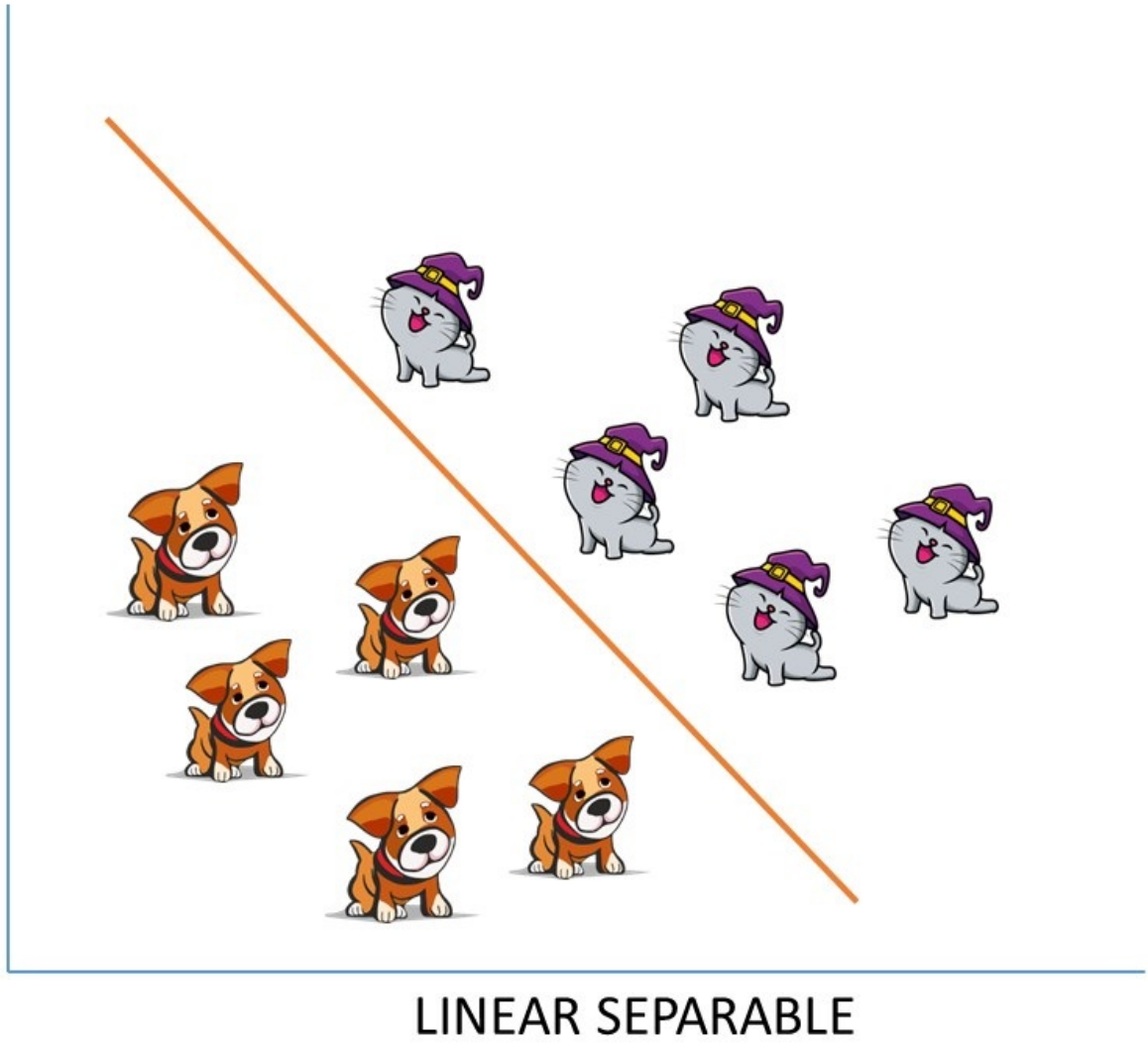
Fornendo alla macchina diverse coppie di input e output (relative ad esempio ad una serie di immagini raffiguranti o meno un determinato oggetto), il Perceptron era in grado di apprendere

autonomamente come associare un input ad un determinato output precedentemente categorizzato, modificando il peso di volta in volta attribuito durante il processo di apprendimento alle relative connessioni tra sensori e circuiti.

Poichè gli esempi forniti per l'addestramento della macchina erano rappresentati da immagini già in precedenza etichettate (dagli operatori umani) come appartenenti a determinate categorie (secondo il modello di apprendimento *supervisionato*), la macchina era in grado di imparare ad associare correttamente le immagini in input alle categorie di appartenenza, stabilendo in questo modo la presenza o meno dell'oggetto all'interno dell'immagine di input.

Una volta terminata la fase di apprendimento, era possibile sottoporre alla rete neurale **nuove immagini** in input, differenti da quelle sottoposte alla macchina in fase di addestramento, e verificare che il Perceptron fosse in grado di classificarle correttamente, assegnandole alle rispettive classi di appartenenza.

Nell'immagine che segue viene mostrato il risultato del processo di apprendimento del Perceptron, con l'attribuzione degli oggetti raffigurati nelle immagini alle diverse classi di appartenenza (in questo caso rappresentate da *cani* e *"stregatti"*):



Classificare con il Perceptron

Come è intuibile dall'immagine, il Perceptron sfrutta un **operatore lineare** (una retta, o un *iperpiano* nel caso di più di due classi di categorizzazione) per discriminare (classificare) i diversi oggetti, assegnandoli alle diverse classi di appartenenza.

Questa modalità caratteristica di classificazione rappresenta anche il **limite concreto e insormontabile** delle capacità predittive del Perceptron, come vedremo tra breve.

I limiti predittivi del Perceptron

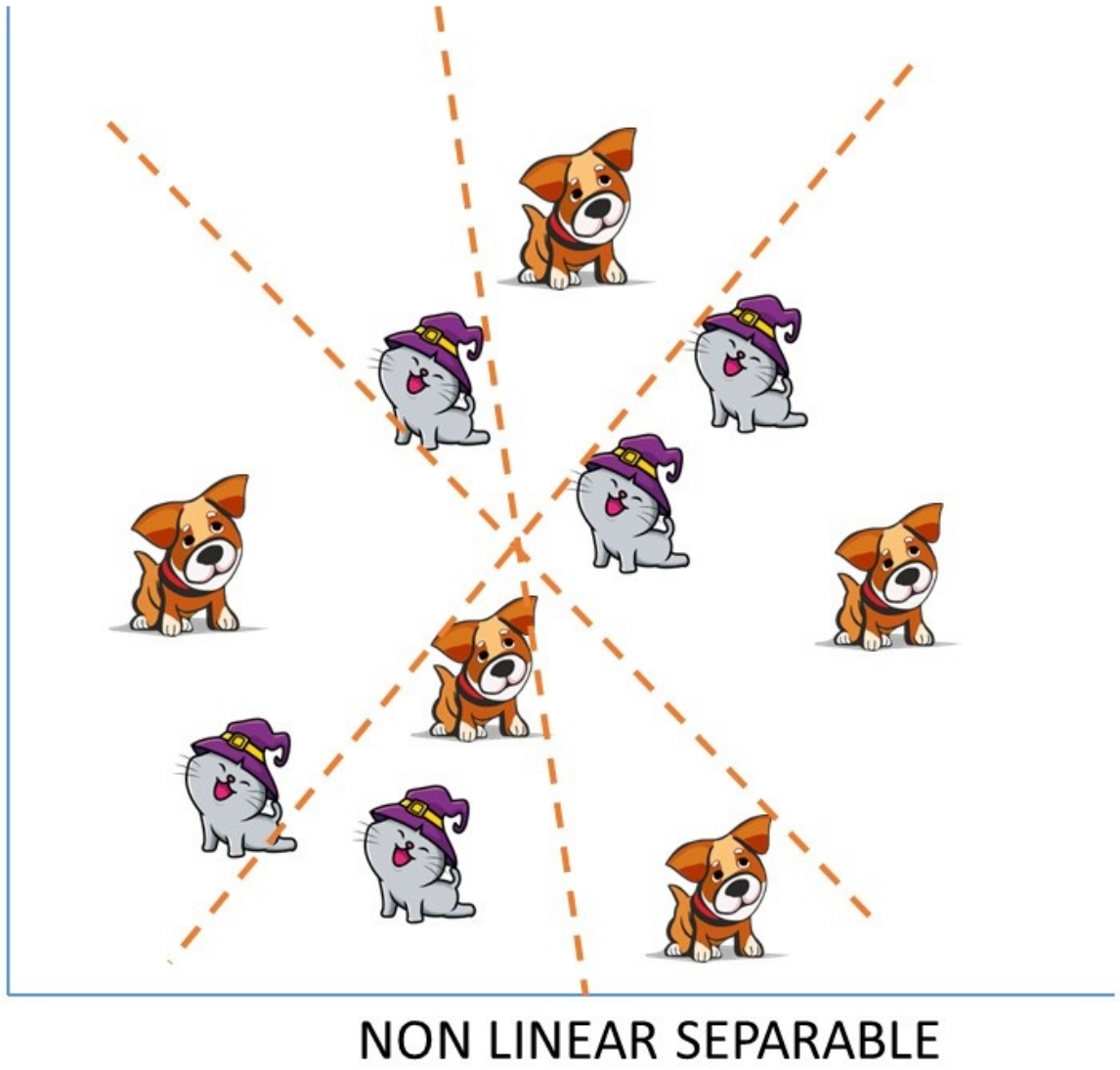
Malgrado il clamore mediatico che seguì alla realizzazione del Perceptron di Rosenblatt, le delusioni dettate dal confronto con la realtà concreta non tardarono ad arrivare.

Fu proprio Marvin Minsky, uno dei partecipanti all'iniziativa inaugurata da McCarthy con il convegno organizzato a Dartmouth nel 1956, a spegnere rapidamente i precoci (ed eccessivamente ottimisti) entusiasmi sul Perceptron, e a dare inizio al primo "inverno" dell'Artificial Intelligence.

Nel paper dal titolo "Perceptrons" scritto insieme a Seymour Papert, Marvin Minsky delineò sia i limiti predittivi che di apprendimento del Perceptron, limiti riconducibili essenzialmente al modello di rete neurale adottato, che si riduce ad **un solo strato** di neuroni artificiali interposto tra i dati di input e lo strato di output.

Tutto ciò implica che il Perceptron è in grado di categorizzare correttamente i dati di input che possono essere **separati linearmente** utilizzando una retta come funzione discriminante.

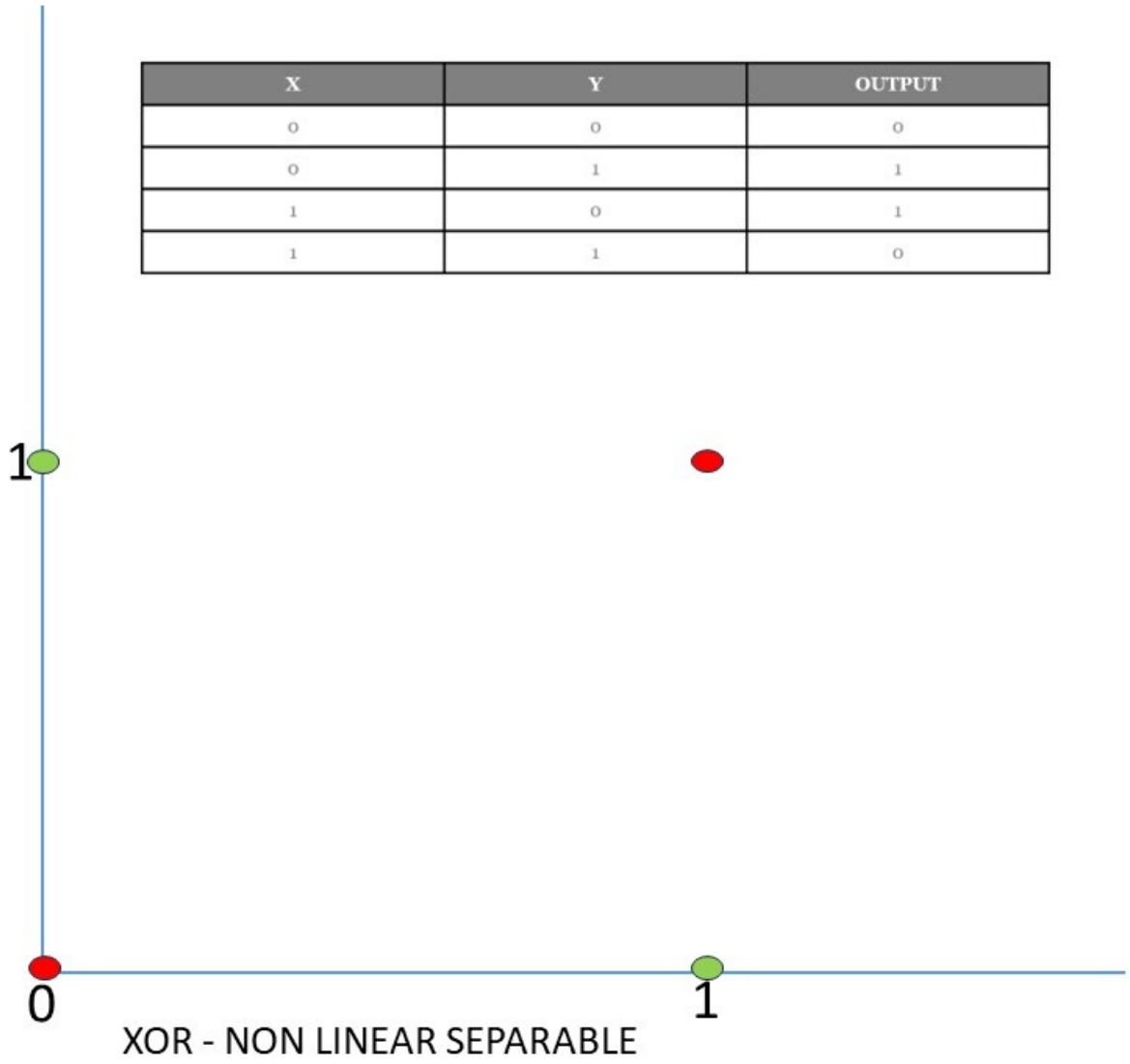
Nel caso in cui questo non fosse vero (nella maggioranza dei casi reali, la funzione che discrimina i diversi oggetti è solitamente più complessa di una semplice retta), durante la fase di apprendimento il Perceptron continuerebbe ad **oscillare indefinitamente** all'interno dei dati di input, nel vano tentativo di individuare la retta che separi correttamente i diversi oggetti da classificare:



Perceptron ed elementi non linearmente separabili

Ad aggravare ulteriormente le cose, vi fu anche la scoperta, sempre da parte di Minsky, della incapacità del Perceptron di apprendere correttamente anche semplici operatori logici come lo **XOR** (or esclusivo), vale a dire la regola decisionale che impone di selezionare “o l’uno, o l’altro” degli elementi, ma “non tutti e due contemporaneamente”.

Nell'immagine che segue, viene riportata la tabella logica associata all'operatore XOR, i cui risultati riportati su uno schema cartesiano, mostrano come sia impossibile individuare una retta che possa individuare un confine di separazione lineare:



Perceptron e operatore logico 'XOR'

Per comprendere intuitivamente i limiti del Perceptron, consideriamo una rete neurale che abbia solo due unità di input, e

che ogni input possa assumere i valori “on” (accesa) oppure “off” (spenta).

Il valore assunto dall’unità di output dipenderà dai valori assunti delle unità di input moltiplicati per i rispettivi pesi.

Ci proponiamo di addestrare la macchina affinchè riconosca se i due valori di input **sono uguali**: essa dovrà pertanto ritornare come risultato dell’unità di output il valore “on” (acceso), nel caso in cui le unità di input siano **entrambe** o “accese” o “spente”.

Se al contrario un input è acceso e l’altro è spento (e viceversa), il risultato di output dovrebbe essere “off”.

Per far sì che questo accada (vale a dire per evitare che l’unità di output mostri “on” in presenza anche di una sola unità di input accesa), occorre aggiustare i relativi **pesi assegnati** alle unità di input in maniera tale che una singola unità di input accesa non sia in grado da sola di attivare il risultato di output su “on”.

Una possibilità sarebbe quella di assegnare alle unità di input dei pesi relativamente bassi, per cui sarebbe necessario l’attivazione di entrambe le unità di input per conseguire il peso complessivo necessario ad attivare l’output.

Peccato però che tale configurazione dei pesi sarebbe adeguata a rappresentare solo 3 su 4 delle possibilità di attivazione: lascierebbe infatti fuori il caso in cui entrambe le unità di input sono “spente”, restituendo un output settato su “off” (quando in realtà esso dovrebbe essere settato su “on”, a indicare che le unità di input si trovano nello stesso stato…)

Per quanto ci sforzassimo di individuare una corretta assegnazione dei pesi, non riusciremmo comunque a soddisfare tutte le esigenze di classificazione contemporaneamente, per il semplice fatto che ciò non è possibile.

Pertanto, il lavoro di Minsky e Papert non solo dimostrava che il Perceptron non era in grado di svolgere alcune semplici operazioni, come decidere se due cose sono uguali oppure no, ma che esso costituiva un modello inadeguato a rappresentare il ragionamento umano, facendo così svanire tutti i sogni di realizzare una “macchina pensante” (e i relativi finanziamenti).

Efficacia predittiva al prezzo della Trasparenza

Come se non bastasse, il meccanismo di apprendimento implicito nel Perceptron realizzava anche il “divorzio” dalla logica binaria, a dispetto di quanto ipotizzato nei modelli di McCulloch e Pitts.

Consentire al Perceptron di adattare autonomamente la rete neurale sottostante mediante il bilanciamento automatico dei pesi associati alle singole connessioni, significava infatti rinunciare ai concetti tipici della logica binaria, stabilita dai corrispondenti operatori booleani (dal nome di George Boole, il primo a introdurre l’algebra della logica binaria).

Al di là del fatto che le unità rappresentative della rete neurale del Perceptron possano assumere gli stati binari “on”/“off”, le regole di apprendimento implementate possono in realtà associare all’attività dei neuroni artificiali **qualsiasi valore**, sulla base dei **diversi pesi** che ad essi sono assegnati.

La maggiore flessibilità così ottenuta richiede un prezzo in termini di **trasparenza e comprensibilità**: in mancanza di un’associazione chiara con gli operatori binari, diventa più difficile associare valori di verità (in senso logico) all’attività delle singole unità componenti la rete neurale.

Rispetto alla trasparenza del modello teorico di McCulloch e Pitts, il Perceptron rischiava di dar vita a un garbuglio incomprensibile: per la prima volta si prospettava il problema delle reti neurali come “black-box”, e il conseguente *trade-off* tra funzionalità e trasparenza, che affligge tutt’oggi l’impiego delle rete neurali artificiali.

Non solo: per la prima volta con il Perceptron veniva mostrato che le reti neurali artificiali potevano effettuare computazioni senza la necessità di rispettare le rigide regole della logica formale.

Tali considerazioni teoriche avevano ricadute importanti anche nell'ambito delle **scienze cognitive**, in modo particolare in relazione all'approccio noto come **"connessionismo"**, già in voga all'epoca, che intendeva spiegare il funzionamento della mente umana mediante le reti neurali artificiali.

Se perfino il Perceptron era in grado di effettuare computazioni senza la necessità di usare proposizioni o operatori logici, si poteva ipotizzare che neanche i neuroni e le connessioni nel cervello avessero bisogno di ricoprire un ruolo definito in termini di logica binaria.

In altri termini, si ipotizzava che anche nel caso del cervello umano, la funzione di una particolare rete fosse distribuita tra i neuroni che la compongono ed emergesse dalle connessioni istituite tra di essi.

Secondo il connessionismo, le informazioni elaborate da una rete neurale, indipendentemente dal substrato impiegato per la sua implementazione (in ossequio al già citato paradigma funzionalista) non sono localizzate in un punto ben preciso della rete, ma devono essere considerate **distribuite tra i nodi** che compongono la rete stessa.

Allo stesso modo, secondo il connessionismo, non ha senso sostenere che una determinata unità di elaborazione all'interno

della rete contenga una determinata informazione, nè che essa svolga un compito o una funzione specifica.

Al contrario, il modello rappresentativo che viene sostenuto è quello della elaborazione delle informazioni **in forma distribuita e in parallelo** (“Parallel Distributed Processing” o modello PDP).

Tale approccio si pone in antitesi con il **cognitivismo**, che pone come principio fondamentale l’assunto che esista una stretta analogia tra mente e computer (come abbiamo visto nella prima parte del testo).

L’approccio connessionista ha quindi molteplici implicazioni “filosofiche”, in quanto ipotizza che l’attività della mente sia distribuita all’interno dell’attività neuronale, e che pertanto non possa essere scomponibile in singoli processi cognitivi, e in tal senso si avvicina al **comportamentismo**, in cui assumono un ruolo centrale le relazioni di stimolo e risposta.

Come vedremo, il connessionismo riceverà nuova linfa con la diffusione delle **reti neurali profonde**, volte a superare i limiti computazionali del Perceptron.

L’Apprendimento Profondo alla riscossa

I limiti del Perceptron, individuati da Minsky e Papert nel loro libro, erano reali; tuttavia, così come il clamore suscitato dal Perceptron era prematuro, allo stesso modo era prematuro dare per “spacciato” l’approccio che tale modello computazionale aveva inaugurato.

Gli stessi Minsky e Papert avevano prefigurato infatti la possibilità che i limiti attuali del Perceptron fossero sostanzialmente attribuibili al fatto di limitarsi a **un solo strato** di neuroni artificiali tra l’input e l’output.

Ma non necessariamente la rete neurale predisposta mediante Perceptron doveva limitarsi a un solo strato: aggiungendo ulteriori strati di neuroni all’interno della rete, era infatti possibile superare i limiti predittivi del Perceptron.

Ad esempio, per risolvere il problema di riconoscere l’uguaglianza di coppie di valori di input, era sufficiente aggiungere al modello originario del Perceptron un ulteriore strato di neuroni tra l’input e l’output.

Questo strato ulteriore sarebbe stato composto da due soli neuroni, uno dei quali si sarebbe attivato nel caso in cui entrambi i valori di input fossero su “on”, e l’altro si sarebbe attivato nel caso contrario, ovvero quando entrambi i valori di input fossero posti su “off”.

In questo modo, il neurone di output sarebbe stato in grado di discriminare i vari stati assunti dai neuroni di input, sulla base delle informazioni ricevute direttamente dai neuroni posti

all’interno dello strato intermedio, riuscendo così a distinguere agevolmente i casi in cui i valori di input fossero posti entrambi su “on” o su “off”.

Con l’introduzione dei Perceptron multistrato si stava prefigurando l’avvento delle **rete neurali artificiali “profonde”**, definite in tal modo proprio in relazione al grado di “profondità” che gli ulteriori strati di neuroni posti tra lo strato di input e quello di output potevano assumere.

Come abbiamo anticipato nella precedente sezione, tale architettura neurale “profonda” avrebbe inoltre dato ulteriore linfa all’approccio connessionista alla mente.

Malgrado Minsky e Papert avessero preconizzato l’introduzione delle reti multistrato di Perceptron, e fossero consapevoli delle loro potenzialità, tuttavia non contribuirono a ripristinare la fiducia perduta verso la realizzazione dei sogni dell’AI.

Ai tempi, i nostri Autori erano ben consapevoli dell’esistenza di un ostacolo considerato insormontabile: come addestrare una rete neurale multistrato.

L’Apprendimento Profondo e la “backpropagation”

In altri termini, non erano note le modalità per addestrare e permettere alle reti neurali profonde l’apprendimento di cui necessitavano.

Mentre nel caso del Perceptron era relativamente semplice impostare le connessioni richieste tra i neuroni di input e output per realizzare l’addestramento della rete, non era affatto chiaro come si dovesse procedere ad impostare le connessioni tra i neuroni nel caso in cui gli strati fossero più di due.

Bisognerà attendere il 1986 e la pubblicazione del paper dal titolo “Learning representations by back-propagating errors” per avere finalmente una soluzione al problema relativo a come addestrare una rete neurale artificiale multistrato.

Tra gli Autori del paper citato vi è anche Geoffrey Hinton, che in futuro verrà soprannominato il “padrino” dell’AI proprio in relazione al contributo fornito alla soluzione del problema dell’apprendimento.

Tale soluzione si basa su un algoritmo noto come “backpropagation”, che consente alla rete di “imparare dai propri errori” predittivi, propagando all’indietro gli errori inoltrandoli agli strati all’interno della rete (ovvero inoltrando gli errori rilevati dagli strati più esterni verso quelli più interni della rete), consentendo così di apportare le dovute correzioni, consistenti essenzialmente nell’aggiustamento dei pesi associati ai singoli neuroni.

Tornando alla rete originaria di Perceptron, costituita da un solo strato di input direttamente connesso allo strato di output, è relativamente semplice apportare le dovute correzioni sulla base degli errori riscontrati nelle predizioni: se un neurone in output dovesse risultare *spento* quando invece era atteso che fosse *acceso* (e viceversa), basterà rinforzare il peso associato alla connessione con il corrispondente neurone di input.

In altri termini, mentre nel caso della rete a strato singolo, la relazione tra le connessioni in input e output è chiara, non altrettanto può dirsi nel caso di una rete con molti livelli intermedi tra input e output.

Era quindi indispensabile individuare una soluzione matematica che risolvesse il problema nella sua forma generale.

Per comprendere l'importanza dell'algoritmo di backpropagation, immaginiamo una rete neurale costituita solo di tre strati: lo strato di input, quello di output, e quello intermedio (in pratica, al modello originario del Perceptron abbiamo aggiunto solo lo strato intermedio).

Rappresentare in forma matematica le relazioni costituite dalle connessioni tra gli **strati contigui** direttamente legati tra di loro (ovvero lo strato di input con quello intermedio, e lo strato intermedio con quello di output) è relativamente semplice.

Si tratta infatti di specificare in un'equazione lo stato dei neuroni direttamente connessi tra di loro, e dei relativi pesi.

Il problema consiste nel trovare la relazione che lega tra di loro le diverse equazioni individuate per i diversi strati contigui, per

definire in maniera completa in che modo le connessioni che partono dallo strato di input influiscono su quello di output, e apportare di conseguenza le dovute correzioni.

L'intuizione centrale dell'algoritmo di backpropagation è stata quella di sfruttare una nota regola del calcolo infinitesimale, nota come **"regola della catena"**, per porre in relazione tra loro le diverse equazioni rappresentative delle connessioni esistenti tra gli strati contigui, e apportare gli opportuni correttivi nei singoli pesi e stati associati ai neuroni dei diversi strati, per consentire le correzioni nelle previsioni restituite dalla rete, permettendo così alla rete stessa di effettuare l'addestramento necessario all'apprendimento.

In questo modo, era finalmente possibile utilizzare reti multistrato e sfruttare l'apprendimento profondo, noto appunto come **"Deep learning"**, per addestrare le reti neurali profonde a compiere i più diversi compiti, conseguendo gli strabilianti risultati che ormai sono noti a tutti, complice inoltre l'accresciuta **capacità computazionale** resa disponibile dal calcolo parallelo (*distributed computing*) che di lì a poco approderà al **Cloud Computing**.

Ma per completare il puzzle mancava ancora un tassello importante…

I Big Data, il carburante delle Reti Neurali Profonde

Si è spesso favoleggiato dei Big Data come del petrolio del nuovo millennio; di una cosa possiamo essere certi: i dati costituiscono il carburante necessario per addestrare le reti neurali profonde.

E di dati ce ne vogliono veramente tanti per poter permettere alle reti neurali profonde di “apprendere”: il tassello mancante del puzzle era appunto rappresentanto dalle moli di dati, e solo l’accumulo di essi realizzato dalla diffusione della rete internet avrebbe permesso alle tecnologie basate sul Deep Learning di potersi affermare con successo.

La disponibilità di enormi quantità di dati ha consentito la possibilità non soltanto di addestrare le reti neurali profonde, ma anche quella di emulare con un elevato grado di verosimiglianza statistica alcune delle funzioni del comportamento umano.

Piuttosto che tentare di replicare le strutture del cervello umano, le reti artificiali profonde si propongono di estrarre le **regolarità statistiche** più significative dai dati disponibili, permettendo così alle macchine di simulare in maniera verosimile il comportamento umano.

Ecco quindi che i servizi di traduzione linguistica come Google Translate, così come gli assistenti virtuali quali Siri e Alexa, sfruttano le reti neurali artificiali profonde per riprodurre con un elevato grado di affidabilità le prestazioni linguistiche di traduzione e conversazione umane.

Sino ad arrivare ad elaborare testi in maniera talmente verosimile da non riuscire a distinguere più tra autore umano e “meccanico”, come nel caso di ChatGPT.

Questi risultati indubbiamente strabilianti, hanno indotto molti (con la complicità degli addetti ai lavori interessati a diffondere tali suggestioni) a credere in qualche forma di “senzienza” raggiunta da tali algoritmi.

Ma in realtà, algoritmi di apprendimento quali il backpropagation non intendono ricreare il comportamento del cervello umano (in questo gli stessi ideatori dell’algoritmo sono stati chiari fin da subito), quanto piuttosto **ottimizzare** l’estrapolazione delle **regolarità statistiche** estratte dalle grosse moli di dati disponibili mediante l’efficace ed efficiente correzione degli errori di predizione dei risultati ottenuti dalla rete nel replicare determinati comportamenti umani (come la scrittura e la comunicazione verbale).

In altri termini, si tratta sempre di **simulazione** di comportamenti, che trova la propria forza persuasiva nella capacità delle reti neurali di individuare le funzioni matematiche più appropriate a rappresentare le regolarità statistiche insite nei dati di apprendimento stessi.

E la potenza predittiva offerta dall’ottimizzazione matematica delle regolarità estratte dalle grandi moli di dati rappresenta anche la causa, come vedremo, di molti dei rischi e dei risultati inattesi prodotti dagli algoritmi.

Per comprendere le ragioni matematiche alla base delle strabilianti capacità predittive offerte dalle reti neurali artificiali profonde, dobbiamo accennare brevemente al teorema di approssimazione universale.

Le Reti Neurali Artificiali e il Teorema di Approssimazione Universale

Senza entrare nei dettagli analitici, dal punto di vista matematico lo scopo a cui mira una rete neurale artificiale è quello di individuare una funzione matematica in grado di fornire l'adeguata mappatura dei dati di input a quelli di output.

Il livello di accuratezza della mappatura fornita da tale funzione matematica varia a seconda della distribuzione dei dati e dell'architettura della rete impiegata.

La funzione matematica individuata dalla rete neurale può essere arbitrariamente complessa.

Il teorema di approssimazione universale ci dice che, a determinate condizioni, le reti neurali hanno la capacità di approssimare qualsiasi funzione matematica, a prescindere dalla sua complessità, e indipendentemente dal numero di attributi in input e output.

In altri termini, le reti neurali sono in grado di individuare la mappatura adeguata a rappresentare le relazioni statistico-matematico che legano tra loro i dati di input e output, al fine di ottenere predizioni caratterizzate da un grado di approssimazione arbitraria.

È in questa capacità di approssimazione universale che consiste la potenza predittiva delle reti neurali profonde, e che permette di realizzare simulazioni dei comportamenti umani caratterizzate da uno strabiliante grado di verosimiglianza.

Ma al contempo tale potenza predittiva costituisce paradossalmente il limite dell'affidabilità pratica, oltre che conoscitiva, delle reti neurali profonde.

---

1. cfr. "Dartmouth workshop" https://en.m.wikipedia.org/wiki/Dartmouth_workshop↩

LO SCIENTISMO POSITIVISTA DELL'ARTIFICIAL IDIOCY

I **limiti pratici** delle reti neurali profonde li vedremo nel prosieguo della trattazione; in questa sede ci soffermiano ad analizzare i **limiti teorici**.

E i limiti teorici sono essenzialmente riconducibili all'approccio metodologico adottato nella realizzazione delle reti neurali e del relativo apprendimento, basato sostanzialmente sul **metodo induttivo** che riafferma nei fatti lo **scientismo positivista** implicito nelle procedure **"data driven"** basate esclusivamente sui dati osservativi.

Abbiamo visto infatti nelle precedenti sezioni come le enormi moli di dati siano indispensabili per realizzare l'addestramento necessario per conseguire l'apprendimento delle reti neurali profonde.

Lo sfruttamento dei big data a tal fine, implicitamente attesta l'adozione dell'epistemologia induttivista insita nell'approccio *data-driven*, ben sintetizzata dal motto tanto caro ai suoi sostenitori: "lasciate che i dati parlino da soli".

L'approccio **induttivista** costituiva non a caso la cifra rappresentativa del **positivismo**, in quanto intendeva scoprire le verità scientifiche ritraendole direttamente ed esclusivamente dai **dati osservativi**.

Tale approccio risuona in maniera evidente in un noto articolo di qualche anno fa a firma di Chris Anderson, che intendeva inaugurare la "rivoluzione" metodologica rappresentata dai **big**

data e dall'approccio **data-driven**, e che nelle intenzioni dell'Autore, avrebbe addirittura reso obsoleto il tradizionale **metodo scientifico** nell'acquisizione di **nuova conoscenza**, come vedremo tra poco.

Approccio data-driven e metodo induttivo

La logica di fondo che ispira l'approccio **data driven** è essenzialmente **induttivista**: si dà per scontato che aumentando il numero di dati di addestramento forniti alla macchina, alla fine essa sarà in grado di scoprire tutto ciò che di rilevante c'è da scoprire del mondo esterno, e tutto il resto, vale a dire tutto ciò che non è presente nei dati osservativi, è per definizione **inesistente** (esattamente come le entità metafisiche per i positivisti).

Ovviamente i fautori dell'approccio **data driven** sono consapevoli della impraticabilità fattuale di mostrare alla macchina **tutti i dati osservativi possibili e immaginabili**, perchè questo significherebbe accumulare quantità tendenzialmente **infinite** di informazioni, con tempi di apprendimento altrettanto infiniti per la macchina, dando luogo inoltre a quella "esplosione combinatoriale" che rende **intrattabile** dal punto di vista **pratico** la gestione dei dati a livello computazionale.

Ma quello che per i fautori dell'approccio data-driven è considerato esaustivo dal punto di vista metodologico (e che pertanto costituisce una "garanzia" dell'affidabilità dell'approccio e dei risultati da esso ottenuti), è che la possibilità di accumulare indefinitivamente dati osservativi sia realizzabile **in linea di principio**.

Anche in questo caso, riecheggia la convinzione positivista che la scienza rappresenti un processo di accumulazione **continuo e**

costante di conoscenza, e che la scienza sia in grado di approssimarsi sempre di più alla conoscenza certa.

Ovviamente si ammette che il processo possa essere lungo (al limite, *infinito*) ma non si pone minimamente in dubbio che le conoscenze così conseguite ci avvicinino progressivamente alla verità scientifica.

L’eredità meccanicistico-deterministica

La mentalità positivista è il frutto da un lato del successo della fisica newtoniana, e dall’altro dall’eredità rappresentata dal modello **meccanicista-determinista** che ne costituisce il retroterra culturale, inaugurato con Cartesio e ben rappresentato ancora nell’ottocento dalla famosa affermazione di Pierre-Simon Laplace: *“noi dobbiamo riguardare il presente stato dell’Universo come l’effetto del suo stato precedente e come la causa di quello che seguirà. Ammesso per un istante che una mente possa tener conto di tutte le forze che animano la natura, assieme alla rispettiva situazione degli esseri che la compongono, se tale mente fosse sufficientemente vasta da poter sottoporre questi dati ad analisi, essa abbraccerebbe nella stessa formula i moti dei corpi più grandi dell’Universo assieme a quelli degli atomi più leggeri. Per essa niente sarebbe incerto e il futuro, così come il passato, sarebbe presente ai suoi occhi*” (Essai philosophique sur les probabilités, 1814).

Per oltre due secoli, la fisica newtoniana ha rappresentato il modello ideale di conoscenza scientifica “certa”, in virtù della capacità predittiva da essa dimostrata nel tempo.

La fisica newtoniana inoltre si dimostrava efficace sia per descrivere i fenomeni quotidiani su distanze “ordinarie”, che per prevedere i movimenti dei corpi celesti.

Essa, pertanto, era ritenuta costituire il “modello definitivo” per la descrizione di tutti i fenomeni fisici, al punto che gli stessi fenomeni studiati dalla nascente branca dell’elettromagnetismo,

venivano ricondotti all'interazione di corpuscoli assoggettati alle stesse leggi di Newton.

Tuttavia, come spesso capita, le suggestioni della *hybris* (l'orgogliosa tracotanza umana, nell'accezione degli antichi greci) sono condannate a sciogliersi come neve al sole, portandosi dietro tutta la protervia e l'arroganza di chi le ha alimentate…

L’irriducibile complessità della realtà

Il modello **meccanicista determinista** franerà da un lato sotto i colpi della nuova scienza della complessità e del caos, inaugurata da quel genio assoluto di Henrì Poincarè (per ironia della sorte anch’egli francese, come Laplace), che per poco si lascerà sfuggire il primato di scopritore della teoria della relatività, primato universalmente riconosciuto ad Einstein.

Lo stesso Einstein, a sua volta, contribuirà a demolire definitivamente l’edificio newtoniano di “certezze” (queste sì basate su entità metafisiche *inosservabili*, secondo la concezione positivista, quali il concetto di forza, azione a distanza, tempo e spazio “assoluti”), ponendo inoltre le basi per quella meccanica quantistica, che egli stesso stentò a considerare una rapprentazione “vera” della realtà, in questo mostrandosi ancora legato al modello teorico precedente che suo malgrado aveva contribuito a smantellare.

E con il modello meccanicista determinista franeranno anche le ambizioni irrealistiche del positivismo: la scienza non è un metodo di acquisizione di conoscenze “certe” stabilite una volta per tutte, nè il suo cammino verso la Verità segue un percorso di approssimazione lineare, ma al contrario, è caratterizzato da un succedersi di paradigmi, più simile ad un percorso accidentato, con arresti e ripartenze da zero.

Malgrado il positivismo fosse stato archiviato come “ferro vecchio” screditato dalla stessa evoluzione della scienza, molte delle sue

suggestioni sono tuttavia ritornate in auge proprio grazie alla **Artificial Idiocy**.

E la ragione è legata alla comunanza di **approccio metodologico** che entrambi condividono: vale a dire, la credenza che la conoscenza possa essere conseguita lasciando che i dati “parlino da soli”.

Anche il positivismo pensava che l’acquisizione delle conoscenze scientifiche si riducesse in ultima analisi nella raccolta dei dati “positivi”, e che le teorie scientifiche trovassero la loro **conferma** nel metodo di **ragionamento induttivo**.

Ciò che non si vede, non esiste…

In altri termini, **solo i dati osservativi** sono considerati necessari e sufficienti per conseguire una conoscenza affidabile, pertanto tutto ciò che non è presente nei dati osservativi, è per definizione inesistente e/o irrilevante.

E tra le cose rese "irrilevanti" dall'approccio *data driven* ci sono le stesse **teorie scientifiche** tradizionali, trattate alla stregua di "narrazioni fantasiose" che tradiscono i *pregiudizi* dei ricercatori, e rappresentano un intralcio alla acquisizione di nuova conoscenza, specie in contesti complessi caratterizzati da flussi informativi esorbitanti (come quelli prodotti da internet e dai servizi della *new economy*).

Al punto che, in un famoso articolo Chris Anderson, l'allora direttore della rivista *Wired* (rivista di culto per i tecnosciovinisti), profetizzò la "Fine della Teoria" 1, sostituita dalla *Big Data Analytics*, che dell'approccio *data driven* costituiva l'espressione più autentica e riuscita.

Anche in questo caso, tuttavia, la realtà non è riducibile alle semplici evidenze osservative, e molte delle relazioni sottostanti ai dati necessitano proprio di una **ricostruzione teorica** per essere portate alla luce, a dispetto di quello che voleva far intendere Anderson…

Senza una teoria, i dati sono muti

> “Il Logos ama nascondersi”
> Eraclito, *Frammento 123*

Tra i tanti argomenti in grado di dimostrare la fallacia dello slogan “lasciate che i dati parlino da soli”, comunemente sbandierato dai fautori dell’approccio *data-driven*, ve n’è uno che reputiamo essere sintomatico e rappresentativo anche dei **bias** che caratterizzano la retorica e la narrazione tipica delle start-up tecnologiche: ci riferiamo alla “fallacia del sopravvissuto”, anche nota come “bias di sopravvivenza”.

Il bias di sopravvivenza consiste nell’errore logico in cui cade chi si concentra esclusivamente o prevalentemente sui casi di successo, trascurando quelli di insuccesso.

Tale bias è riconoscibile anche nella retorica che esalta i successi finanziari delle innovazioni di successo delle **start-up** tecnologiche, focalizzandosi di conseguenza sulle iniziative imprenditoriali che hanno superato il processo di selezione rappresentato dal mercato, trascurando tuttavia quelle che non ci sono riuscite, che spesso rappresentano la maggioranza (quando non la quasi totalità) dei casi.

Occorre ricordare infatti che la gran parte delle iniziative imprenditoriali tecnologiche, e le relative innovazioni da esse proposte, finisce ad ingrossare il cosiddetto “cimitero dei

fallimenti", ovvero sono relegate rapidamente nel dimenticatoio e vengono (erroneamente) espunte dai report di settore.

Pertanto, la fallacia del sopravvissuto può indurre a conclusioni errate a causa della incompletezza informativa che essa comporta.

L'incompletezza informativa coinvolge anche l'approccio data-driven, non solo perchè esso, come abbiamo visto nei paragrafi precedenti, dà per scontato che sia sempre possibile acquisire i dati necessari a risolvere un determinato problema (supponendo addirittura che i dati e le informazioni reperibili sulla rete rappresentino l'universo dei dati campionari inteso in senso statistico).

Ma anche e soprattutto perchè non tiene in considerazione l'importanza delle **informazioni implicite** non immediatamente rilevabili dai dati osservativi, che necessitano di un modello teorico adeguato (che spesso implica un approccio "creativo") che le faccia emergere nella loro evidenza.

L'esempio che proponiamo di seguito intende mostrare un caso da manuale, che mostra in tutta evidenza i limiti suddetti.

E l’aereo tornò solo…

Tra i lettori “diversamente giovani” che leggendo il titolo di questo paragrafo avranno riconosciuto la citazione alla nota canzone resa celebre dalla parodia di Renato Carosone 2, ci sarà magari anche qualcuno che avrà colto il riferimento al famoso caso degli aerei alleati “sopravvissuti” durante la Seconda Guerra Mondiale.

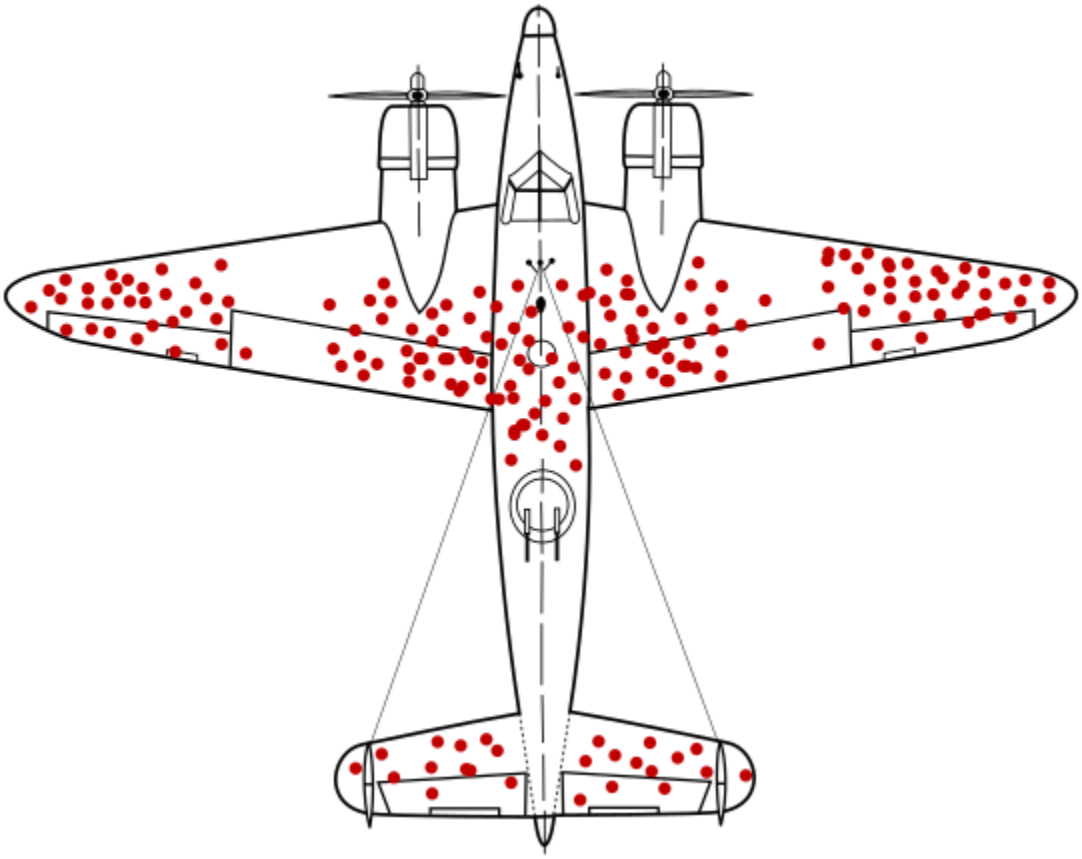
In breve, il caso è riassumibile così: a seguito delle notevoli perdite subìte dalla Royal Air Force inglese e in generale dall’aviazione alleata, causate dalla Luftwaffe e dalla contraerea tedesca, i vertici militari alleati si prefissero l’obiettivo di minimizzare le future perdite a danno della flotta residua, rinforzando i velivoli superstiti nelle parti considerate più deboli, individuate sulla base delle evidenze osservative ottenute analizzando lo stato degli aerei “sopravvissuti” alla battaglia.

Per questo compito, fu dato incarico a Abraham Wald, all’epoca considerato uno dei massimi esperti di statistica.

Il compito che attendeva Wald era tutt’altro che semplice, e comportava dei rischi rilevanti: era necessario infatti non soltanto individuare correttamente le parti degli aerei che necessitavano effettivamente di essere rinforzate, ma occorreva anche minimizzare tali interventi di riforzo, per non appesantire eccessivamente il velivolo, rendendolo di conseguenza meno agile in battaglia.

A complicare ulteriormente il compito, vi era poi il fatto che le evidenze osservative rilevate sugli aerei superstiti sembravano

suggerire la necessità di rinforzare le parti colpite dai proiettili, come appare evidente dall'immagine sottostante:



Fallacia del sopravvissuto (survivorship bias) - Image credits: Wikipedia

È qui che il genio di Wald fa la sua apparizione: adottando un approccio **controfattuale**, ovvero formulando **ipotesi teoriche** come è d'uso nella migliore tradizione del **metodo scientifico** e del **ragionamento ipotetico** che su di esso si basa (definito da C. S. Peirce come **"ragionamento abduttivo"** 3), Wald fu in grado di scoprire quali fossero in realtà le aree dei velivoli meritevoli di essere rinforzate.

L'ipotesi controfattuale avanzata da Wald si traduce in un ragionamento "in negativo":

- se i dati osservativi disponibili riguardano soltanto gli aerei che sono stati in grado di rientrare alla base, vuol dire che i danni da essi subiti **non sono così gravi** da impedirne la sopravvivenza;
- pertanto, **da questa ipotesi** segue che gli aerei abbattuti (di cui non si dispongono le evidenze osservative, in quanto non sono potuti rientrare alla base) hanno subìto danni nelle **parti non evidenziate** dalla mappa dei proiettili evidenziata sugli scafi degli aerei superstiti;
- la conclusione che se ne deve trarre (sulla base del **ragionamento abduttivo**) è che le parti da rinforzare siano quelle che risultano essere **non colpite dai proiettili**, evidenziate "per differenza" analizzando i dati osservativi rilevati sugli scafi degli aerei superstiti;

In altri termini, Wald cercò di **inferire le possibili cause** dell'abbattimento degli aerei non sopravvissuti, senza avere evidenze osservative dirette, e anzi, scartando le evidenze osservative disponibili.

Per conseguire un risultato simile, occorre fare ricorso alla facoltà di **immaginare scenari alternativi** (*controfattuali*), che ci consentono di **formulare ipotesi** sulle **cause** (retrostanti o sottostanti) che possano aver dato luogo ai fenomeni che osserviamo.

Tutto quello che l'approccio induttivo non è in grado di conseguire: anzi, al contrario, rischia di fare la fine del "tacchino induttivista" di Bertrand Russell.

Quando i dati osservativi ingannano

Quello del “tacchino induttivista” di Russell [4] è un esempio lampante dei **limiti costitutivi** dell’approccio **induttivo** alla **conoscenza**: malgrado l’apparenza di oggettività e di “certezza” che sembra suggerire, in realtà esso può indurre a **conclusioni fatalmente errate**, che risultano esiziali, come le previsioni formulate dal tacchino alla vigilia di Natale…

Paradossalmente, l’osservazione ripetuta condotta su 358 giorni all’anno (computati da inizio anno fino al giorno della vigilia di Natale) durante i quali il tacchino osserva il fattore portargli il cibo, raggiunge la sua **massima verosimiglianza** statistica proprio il giorno della vigilia di Natale: in altri termini, le “evidenze” raccolte dal tacchino fino a quel momento, supportano al massimo grado l’aspettativa di vedere il fattore portargli il cibo.

L’ipotesi controfattuale che il fattore alimenti il tacchino con l’obiettivo di trasformarlo a sua volta in cibo, non emerge in alcun modo dai dati osservativi a disposizione del povero volatile…

Mentre le previsioni formulate sulla base dell’approccio induttivo si limitano ai dati osservativi disponibili, per converso, l’approccio ipotetico (*abduttivo*, nell’accezione di Peirce) ha il vantaggio di **espandere la conoscenza disponibile** andando oltre i meri dati osservativi, pur accettando la possibilità che le ipotesi formulate possano rivelarsi errate, ma proprio per questo rivedibili, formulando ipotesi alternative.

Queste considerazioni teoriche sulle differenti inferenze predittive che possono essere ottenute sulla base dei diversi approcci

adottati, sono suscettibili di avere conseguenze pratiche importanti, che a loro volta evidenziano i **limiti sostanziali** (e non meramente "accidentali") dell'applicazione dell'intelligenza artificiale alla realtà concreta, dando luogo alla **Artificial Idiocy**, come vedremo nei paragrafi che seguono.

---

1. cfr. "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete" https://www.wired.com/2008/06/pb-theory/ ↩
2. Renato Carosone, "E la barca tornò sola" https://www.youtube.com/watch?v=HOj8RYI5jR0↩
3. cfr. https://it.m.wikipedia.org/wiki/Charles_Sanders_Peirce↩
4. cfr. https://it.m.wikipedia.org/wiki/Tacchino_induttivista↩

I VEICOLI A GUIDA AUTONOMA COME ICONA DEL PROGRESSO

Come primo caso di studio paradigmatico della *Artificial Idiocy*, intendiamo proporre quello relativo ai veicoli "a guida autonoma".

Ci riferiamo più precisamente alle **self-driving cars**, ovvero alle automobili che si pretende siano in grado di guidare e orientarsi nel traffico cittadino quotidiano in maniera **completamente autonoma**, senza necessità di intervento del guidatore umano.

Ne analizziamo le caratteristiche in quanto esse rappresentano un "caso da manuale" in cui gli **incentivi economici** della narrazione che di esse ne danno le case produttrici, si allineano con quelli dei media che ne parlano enfatizzandone i pregi: entrambi hanno infatti da guadagnare dal clamore mediatico che inevitabilmente contraddistingue una innovazione come quella delle automobili che si guidano "da sole".

Del resto, la narrazione associata alle automobili a guida autonoma fa leva su un immaginario fantascientifico a cui il pubblico è stato ormai preparato da tempo, anche grazie ai romanzi e ai film di fantascienza, che da sempre presentano come immagine iconica rappresentativa del progresso tecnologico la capacità delle macchine di esibire una propria **"autonomia" decisionale**, che non sia più strettamente associata a regole rigide predefinite, ma sia frutto di un processo di "apprendimento".

A supporto delle suggestioni fantascientifiche vengono poi addotte argomentazioni attinenti alla supposta **maggiore sicurezza** della guida autonoma, rispetto ai guidatori umani, anche in questo caso

ricorrendo alla consueta retorica tecno-sciovinista basata sulle pretese superiori capacità decisionali degli algoritmi.

A sostegno dell'asserita maggiore sicurezza delle automobili a guida autonoma viene infatti sostenuto che, a differenza dei guidatori umani, gli algoritmi non sono soggetti agli errori derivanti oltre che dalla imperizia nella guida, anche da cause tipicamente umane, quali la stanchezza, la distrazione, o il consumo di sostanze che alterano la lucidità e la responsività agli eventi esterni, quali alcol e droghe.

Ma davvero le automobili a guida autonoma sono più sicure?

E su quale base è possibile stabilire un confronto affidabile con i guidatori umani?

Come vedremo tra breve, gran parte delle strabilianti capacità predittive degli algoritmi di guida autonoma sono in realtà destituite di fondamenti concreti, essendo frutto della retorica e della narrazione tipiche della *Artificial Idiocy*.

Il traffico cittadino e la sua connaturata imprevedibilità

Il traffico cittadino costituisce uno tra i più realistici **contesti dinamici** nei quali sperimentare in concreto la reale **autonomia decisionale** delle macchine.

Esso è infatti caratterizzato da **imprevedibilità** degli eventi esterni e dalla **complessità** delle interazioni tra i diversi agenti coinvolti (guidatori, pedoni, forze di polizia, servizio pubblico, ecc.) e le mutevoli condizioni dell'ambiente esterno (condizioni metereologiche, visibilità, stato del manto stradale, segnaletica, ecc.)

A questo va aggiunta la **incompletezza informativa** associata alle regole di comportamento e *convenzioni non verbali* che costituiscono la **conoscenza implicita** su cui si basano gli agenti umani nel prendere le loro decisioni nel traffico cittadino, come ad esempio lo *sguardo di intesa* tra guidatori umani in prossimità di un incrocio, al fine di stabilire a chi assegnare la precedenza sulla base della situazione reale (la decisione relativa alla precedenza da essegnare all'uno o all'altro, viene concordata tra i guidatori anche prescindendo in tutto o in parte dalle consuete regole del codice della strada, a causa ad esempio della presenza di ostacoli inattesi sulla carregiata ecc.)

Nel loro insieme, i diversi fattori rappresentativi del traffico cittadino contribuiscono a creare un ambiente quanto più distante dal quel **contesto deterministico** che sarebbe invece auspicabile,

affinchè i modelli predittivi basati su procedure statistiche possano spiegare appieno le loro capacità "predittive".

Per quanti dati storici si possano analizzare, un *contesto dinamico* è caratterizzato da una **imprevedibilità fondamentale**, derivante dalla complessa interazione tra fattori imponderabili.

Senza fare ricorso alle note osservazioni sui fenomeni caotici effettuate da Edward N. Lorenz in ambito metereologico, che hanno dato origine alla "Teoria del Caos" 1, è sufficiente sottolineare come **eventi rari** statisticamente, possano tuttavia avere un **impatto rilevante** in termini di sicurezza.

Non intendiamo riferirci soltanto ai cosiddetti **"cigni neri"**, resi noti del N. Taleb nel suo famoso libro *"Il cigno nero. Come l'improbabile governa la nostra vita"*, in cui l'Autore descrive l'impatto di fattori imponderabili che hanno dato origine alla crisi finanziaria globale del 2007-2008.

I fattori imponderabili alla base dei *cigni neri* sono solitamente riconducibili all'**incertezza**, vale a dire sono legati alla **mancanza** di un **modello predittivo** adeguato a descriverli, o alla inadeguatezza delle teorie scientifiche disponibili.

Non a caso, l'esempio della scoperta dell'esistenza dei cigni neri, avvenuta solo nel 1697 nel continente australiano ad opera dell'esploratore olandese Willem de Vlamingh, è utilizzato da K. R. Popper per sottolineare l'insufficienza del **metodo induttivo** per conseguire l'autentica conoscenza scientifica, contrapponendo ad esso il **falsificazionismo** (per quante osservazioni possa fare,

nessuna mi darà ragione in maniera conclusiva, ma anche una sola osservazione contraria potrà dirmi che ho torto).

Anche in presenza di modelli descrittivi affidabili per la valutazione dei rischi, è sempre possibile che **eventi rari** vengano interpretati in maniera errata dagli algoritmi, a causa della loro assenza nei dataset con cui è stato effettuato l'addestramento del modello.

Affidarsi quindi esclusivamente sulle osservazioni statistiche, rischia di rivelarsi controproducente sotto diversi punti di vista.

Per non parlare poi del fatto che spesso le "evidenze" statistiche vengono opportunamente addomesticate per confermare tesi preconcette...

Le statistiche sugli incidenti e il gioco delle tre carte

Come detto in precedenza, una delle convinzioni più diffuse riguardanti l'Intelligenza Artificiale è quella che sostiene che gli algoritmi sarebbero meno soggetti ad errori, oltre a essere immuni alle idiosincrasie tipiche degli operatori umani.

Pertanto, visto l'incidenza letale degli errori umani in contesti quale la guida nel traffico cittadino, una delle ragioni solitamente addotte a favore della introduzione delle automobili a guida autonoma, si basa appunto su questi argomenti.

Non fosse che l'evidenza dei fatti testimonia in realtà del contrario.

Uno dei primi incidenti mortali che ha coinvolto un veicolo a guida autonoma risale al marzo del 2018, ed è relativo all'investimento di un passante da parte di una autovettura a guida (semi) autonoma di Uber.

Da una ricostruzione della dinamica dell'incidente, le colpe sarebbero riconducibili da un lato ai comportamenti imprudenti tenuti dalla vittima, e dall'altro alla scarsa tempestività del co-pilota nel prendere il comando della vettura (nel caso in questione, si trattava infatti di un veicolo a guida *semi-autonoma*, ovvero che prevedeva la possibilità dell'intervento del guidatore umano).

Tale ricostruzione tuttavia sembra sottostimare le conseguenze negative della possibile presenza di "bachi" nel software, oltrechè dei limiti metodologici di fondo che caratterizzano gli algoritmi utilizzati nella guida autonoma.

Malgrado ciò, i sostenitori *a prescindere* della superiorità algoritmica dei veicoli a guida autonoma, hanno continuato imperterriti a dare per scontata la loro maggiore sicurezza rispetto ai guidatori umani.

Peraltro, facendo un uso piuttosto discutibile delle **statistiche disponibili** sugli incidenti stradali: poichè i veicoli che viaggiano su strada sono stati finora praticamente sempre guidati da esseri umani, è evidente che la maggiore incidenza dei sinistri debba essere attribuita ai guidatori umani.

Il fatto che statisticamente si riscontrino meno incidenti in cui sono coinvolti i veicoli a guida autonoma, è verosimilmente riconducibile al **minor numero di ore guidate** dalle self-driving cars; non per questo soltanto, si possono quindi trarre conclusioni sulla loro maggiore sicurezza.

Per fare un confronto attendibile tra le diverse statistiche, sarebbe quantomeno necessario rendere **omogenei i dati** oggetto di confronto; altrimenti il confronto assomiglia più al famoso “gioco delle tre carte” (o piuttosto alle scommesse vinte in partenza, del tipo “testa vinco io, croce perdi tu”…)

Peraltro, anche soltanto valutando i pochi dati statistici disponibili sugli incidenti che vedono coinvolti i veicoli a guida autonoma, emergono elementi che destano una certa preoccupazione.

Davvero le automobili a guida autonoma sono più sicure?

A dispetto della narrazione trionfalistica portata avanti dalle case produttrici e dai loro CEO (Elon Musk, *patron* di Tesla primo tra tutti), le recenti statistiche sulla sicurezza delle automobili a guida autonoma appaiono impietose.

Tesla, uno dei principali produttori di tali veicoli, ha conseguito il non auspicato primato di risultare in cima alla lista della maggior parte degli incidenti associati all'impiego delle funzioni avanzate di assistenza alla guida (ADAS) - nello specifico le funzioni del pilota automatico - nel primo rapporto NHTSA sulle "Prestazioni di sicurezza delle tecnologie avanzate dei veicoli" 2.

Il report è suddiviso in due sezioni: una per i sistemi avanzati di assistenza alla guida SAE Level 2 e una per i sistemi di guida automatizzata SAE Level 3-5.

I sistemi con cruise control sensibile al traffico e funzioni di mantenimento della corsia rientrano nel livello 2.

Tesla Autopilot rientra in quella categoria e, secondo il rapporto, è in testa per numero di incidenti con un ampio margine.

Occorre dire per completezza che il report non è adattato al numero di veicoli su strada che viaggiano con ADAS, né al numero di miglia percorse con ADAS,

È ragionevole ritenere tuttavia che i veicoli in circolazione dotati di funzionalità ADAS siano riconducibili maggiormente a Tesla, in virtù del fatto che questo produttore ha incluso per anni il servizio Autopilot gratuitamente su tutti i veicoli, laddove la maggior parte

delle altre case automobilistiche faceva pagare un contributo specifico per le funzionalità ADAS.

Secondo una recente analisi del Washington Post condotta sui dati della National Highway Traffic Safety Administration, il sistema di assistenza alla guida di Tesla, noto come Autopilot, risulterebbe coinvolto in molti più incidenti di quanto riportato in precedenza (736 incidenti, e 17 vittime 3).

Al di là dei numeri (che anche nel caso delle automobili a guida autonoma dovranno essere normalizzati sulla base delle serie storiche disponibili di volta in volta), quello che più colpisce dal rapporto stilato dal Washington Post è che molti degli incidenti occorsi (come ad esempio quello di imbattersi in una motocicletta che il sistema automatico apparentemente non è stato in grado di riconoscere) sarebbero stati verosimilmente evitati da un guidatore umano.

Da quanto detto, emerge pertanto che il tipo di errori a cui sono soggetti gli algoritmi delle automobili a guida autonoma è del tutto inedito e inaspettato (altro motivo per il quale fare confronti con gli umani è quantomeno improprio...)

Tali errori fanno ritenere che anche gli algoritmi possano avere degli “abbagli” e comportarsi alla stregua di “ubriachi”, senza necessità di consumare alcolici...

Anche le automobili a guida autonoma si ubriacano

Tra i tanti “abbagli allucinatori” che hanno destato stupore, vi è senza dubbio quelli relativi alla facilità con cui è possibile ingannare gli algoritmi delle automobili a guida autonoma, inducendole a tenere “comportamenti” inattesi, che rappresentano rischi altrettanto inauditi per la sicurezza stradale.

Come mostrato da alcuni ricercatori dell’Università di Washington (vedi l’articolo “Researchers Fool Self-Driving Cars With Stickers on Street Signs” 4) è possibile sfruttare un espediente molto semplice per indurre le auto a guida autonoma a identificare erroneamente i segnali stradali.

I ricercatori hanno scoperto che posizionando strategicamente degli adesivi sui segnali stradali è possibile ingannare il software di elaborazione delle immagini nelle auto a guida autonoma.

Alcuni adesivi attaccati a un segnale di stop hanno fatto sì che i sensori della vettura lo identificassero erroneamente come segnale di limite di velocità.

Il problema risiede nel fatto che la maggior parte dei sistemi di guida autonoma confronta ciò che l’auto “vede” attraverso le sue telecamere con le immagini memorizzate in precedenza.

Pertanto, anche lievi modifiche afferenti all’aspetto esteriore del segnale stradale (che per un operatore umano non comporterebbero alterazioni sostanziali) possono causare errori di predizione dell’algoritmo.

In un altro caso, il segnale stradale di “svolta a destra” è stato offuscato con uno sfondo pixelato grigio-bianco. La vettura in conseguenza delle modifiche apportate (sostanzialmente irrilevanti agli occhi umani, ma non per i sensori) ha interpretato l’oggetto come segnale di “stop”.

Tali errate interpretazioni, oltre a poter essere sfruttate agevolmente da malintenzionati dando luogo a compromissioni di sicurezza, testimoniano la fragilità degli algoritmi anche a minime variazioni nell’ambiente circostante.

Ovviamente quello dell’inadeguatezza della tecnologia attuale nel rispettare le promesse della guida autonoma completa non è un problema soltanto di Tesla (per quanto essa si sia distinta fin da subito per il clamore che ha alimentato a riguardo), ma coinvolge in generale tutto il settore dei veicoli a guida autonoma: fin tanto che la tecnologia non sarà in grado di implementare algoritmi in grado di interagire sinergicamente con i comportamenti degli agenti umani, difficilmente si potrà beneficiare di una autentica e completa autonomia di guida.

E l’obiettivo, per quanto in astratto conseguibile, è in concreto tutt’altro che “dietro l’angolo”, come vorrebbero far credere i produttori e i tecno-sciovinisti…

Gli algoritmi “non lo fanno meglio”

Comprendere i limiti degli algoritmi che presiedono alla guida autonoma degli autoveicoli ci consente di svelare i limiti associati al corrispondente processo decisionale automatizzato, al fine di prevenire i potenziali rischi per l’incolumità fisica.

Tali limiti sono riconducibili essenzialmente alla strategia “guidata dai dati” (*data-driven*) che ispira i processi decisionali automatizzati.

La disponibilità crescente di grandi moli di dati (e la capacità di elaborarli) ha permesso lo sviluppo di applicazioni fino a poco tempo fa inimmaginabili.

Gli algoritmi si sono evoluti al punto di riuscire a sfruttare la crescente quantità di dati disponibili, rendendo possibili servizi avanzati che vanno dalla ricerca personalizzata dei contenuti sulla base delle preferenze degli utenti, alla individuazione delle località di maggior interesse sfruttando la geolocalizzazione, alla traduzione automatica dei testi in diverse lingue.

Per quanto il livello di precisione dei risultati ottenuti tramite tali applicazioni “intelligenti” cresca di giorno in giorno, esse tuttavia non sempre forniscono risultati ottimali.

Per di più, **differenti livelli di precisione** possono comportare **conseguenze diverse** a seconda dei contesti di utilizzo di tali applicazioni e servizi.

Pensiamo ai software di **traduzione automatica** dei testi: in molti casi, i risultati delle traduzioni automatiche possono apparire strabilianti, in altri casi sono invece imprecisi, specie in presenza di contenuti ambigui o di difficile interpretazione per la macchina.

In questi casi, a supplire alle risposte inadeguate fornite dagli algoritmi c'è la capacità di comprensione dei testi tipicamente umana, che cerca di interpretare correttamente i risultati imprecisi o non appropriati offerti dalla macchina.

Stesso dicasi per i risultati ottenuti dai motori di ricerca, che spesso forniscono risposte fuorvianti o non pertinenti.

Pertanto, non è affatto di secondaria importanza che nel progettare i servizi e le applicazioni "intelligenti" ci si rivolga a un utente finale "umano".

La rilevanza del “man in the loop”

Il ruolo svolto dell’utente fruitore del servizio è infatti di fondamentale importanza nel raffinare ulteriormente la ricerca, al fine di ottenere risultati in linea con le aspettative, oppure nel filtrare adeguatamente i risultati ottenuti eliminando quelli non considerati appropriati.

In questo modo, anche se inconsapevolmente, gli operatori umani contribuiscono a colmare le lacune interpretative degli algoritmi (a dispetto delle pretese dei fornitori di tecnologia di voler estromettere l’intervento umano dal “loop”, vale a dire dal ciclo produttivo).

Alla luce di quanto detto, il grado di precisione ottenuto dalle previsioni degli algoritmi *data-driven* avrà un impatto differente a seconda dei diversi contesti.

Nel caso delle traduzioni automatiche, anche un livello di precisione insoddisfacente potrà risultare comunque di qualche utilità per gli scopi che l’utente si prefigge, come ad esempio ottenere una traduzione sufficiente precisa per farsi comprendere da un eventuale interlocutore straniero (anch’esso in grado di colmare le imprecisioni della traduzione facendo ricorso alle proprie competenze linguistiche, oltre che semantiche).

Questo può non essere altrettanto vero quando dalla precisione dei risultati può dipendere l’incolumità fisica delle persone, come nel caso degli errori di interpretazione dei segnali stradali da parte delle automobili a guida autonoma, o l’incapacità di

“comprendere” adeguatamente il contesto di riferimento in cui la macchina è chiamata a prendere delle “decisioni”.

Se le automobili a guida autonoma bloccano il traffico cittadino

A causa dell'incapacità di comprendere adeguatamente il comportamento umano (non soltanto dei guidatori, ma anche dei pedoni), le auto a guida autonoma potrebbero peggiorare il traffico cittadino anzichè agevolarlo, secondo i risultati di un recente studio condotto dall'Università di Copenhagen 5.

Secondo lo studio citato, le auto a guida autonoma contribuirebbero alla congestione del traffico e potrebbero essere potenzialmente pericolose proprio a causa della loro incapacità di comprendere il comportamento umano.

I veicoli a guida autonoma, spesso promossi come il trasporto del futuro, fanno fatica a interpretare i **sottili segnali sociali umani** che informano le decisioni di guida.

Un esempio chiave di questo problema si concentra sulla decisione se cedere il passo o procedere nel traffico, una decisione che gli esseri umani in genere prendono **in modo rapido e intuitivo**.

Tuttavia, le auto a guida autonoma non riescono costantemente a interpretare il comportamento umano nel traffico, e di conseguenza, a causa delle loro reazioni inadeguate al contesto, possono portare alla congestione del traffico.

A detta del professor Barry Brown, del Dipartimento di informatica di Copenaghen, che ha condotto studi sull'evoluzione del comportamento delle auto a guida autonoma su strada negli ultimi cinque anni, "la capacità di orientarsi nel traffico si basa su

molto più delle regole del traffico. Le interazioni sociali, compreso il **linguaggio del corpo**, svolgono un ruolo importante quando ci segnaliamo a vicenda nel traffico. È qui che la programmazione delle auto a guida autonoma è ancora insufficiente. Ecco perché è difficile per loro capire costantemente quando fermarsi e quando qualcuno si ferma per loro, il che può essere sia fastidioso che pericoloso."

A confermare quanto detto, possiamo citare quanto accade per le strade di San Francisco, a seguito dell'autorizzazione concessa alla circolazione dei taxi a guida autonoma, causando ogni tipo di problema, dal blocco del traffico, alla guida sui marciapiedi, lasciando inoltre la città nell'impossibilità di fermarli 6.

L'aspetto maggiormente avvilente di tutto questo, è che non rappresenta in realtà una sorpresa: era infatti ampiamente noto già da tempo agli "addetti ai lavori" che l'attuale tecnologia di guida autonoma fosse inadeguata a gestire la dinamicita e imprevedibilità di un contesto complesso come il traffico cittadino (al punto che la stessa Tesla si è vista costretta a richiamare 362.000 veicoli 7).

Malgrado questo, i produttori di veicoli a guida autonoma hanno fatto leva su tutte le loro capacità (sia finanziarie, che di marketing) di convincimento per far credere esattamente il contrario.

E tra le leve più (ab)usate, c'è sicuramente quella che si appoggia sulle malintese capacità predittive degli algoritmi.

Se gli algoritmi danno la risposta giusta per i motivi sbagliati

Uno degli aspetti più controversi che caratterizza gli algoritmi è costituito dalla irragionevole affidabilità ad essi attribuita, determinata peraltro dall'innegabile efficacia predittiva che essi sembrano mostrare.

In altri termini: la fiducia che viene riconosciuta alle capacità predittive degli algoritmi è diretta conseguenza dell'efficacia predittiva ritraibile dai dati stessi.

Proprio l'efficacia predittiva dei dati ha ispirato il recente paradigma decisionale "data-driven" citato in precedenza, che come abbiamo detto è alla base delle procedure algoritmiche di apprendimento automatizzato.

Tuttavia, non sempre tale efficacia predittiva ci autorizza a fidarci ragionevolmente dei risultati ottenuti dalle procedure automatizzate "guidate dai dati".

Come vedremo infatti tra breve, **efficacia predittiva** non sempre è sinonimo di **affidabilità**.

Per rendersene conto, basta prendere in considerazione i risultati ottenuti utilizzando uno degli strumenti più diffusi che sfrutta la potenza predittiva degli algoritmi di apprendimento automatizzato in abbinamento alle grandi moli di dati: i traduttori automatici.

È innegabile che negli ultimi anni tali strumenti siano migliorati in maniera impressionante, e le traduzioni automatiche di un testo da e verso differenti lingue, ottenute mediante tools di intelligenza

artificiale come Google Translate, costituiscono in definitiva delle traduzioni **“per lo più”** affidabili.

Tuttavia è in quel “per lo più” che si nasconde il diavolo: anche se non affidabili al cento per cento, si suppone che tali traduzioni vengano interpretate da esseri umani “senzienti”, vale a dire in grado di comprendere il “senso” di tali traduzioni.

L’affidabilità statistica di tali traduzioni, in altri termini, rimanda ad operatori umani il compito di risolvere i **casi limite** più inconsueti, vale a dire quelli caratterizzati da maggiore ambiguità.

In realtà, per la maggioranza dei casi, i traduttori automatici si affidano a regole meccaniche di derivazione delle regole di associazione tra differenti raggruppamenti di parole, inferite sulla base delle rispettive **probabilità** stimate.

In altre parole, il meccanismo di traduzione automatica ci fornisce *“per lo più”* la **traduzione corretta**, facendo tuttavia ricorso alle **ragioni sbagliate**.

In altri termini, il traduttore automatico funziona non perchè esso sia in grado effettivamente di comprendere il **senso del testo**, ma perchè **statisticamente** è possibile istituire una relazione biunivoca sufficientemente affidabile tra le diverse rappresentazioni linguistiche dello stesso testo.

Salvo ovviamente nei **“casi limite”**, come ad es. nel caso delle metafore, delle espressioni ambigue o polisemiche, ecc.

Il punto critico è che, mentre nel caso delle traduzioni automatiche eventuali imprecisioni non sono suscettibili di mettere a repentaglio l’incolumità fisica degli esseri umani (anche se ne

potremmo discutere…), ciò non è altrettanto vero nel caso degli errori commessi dagli algoritmi dei veicoli a guida autonoma, come abbiamo visto.

Pertanto non è possibile trarre la conclusione di **maggiore affidabilità** di una procedura decisionale automatizzata facendo ricorso esclusivamente alla sua **efficacia predittiva**.

Non soltanto per gli algoritmi delle automobili a guida autonoma, ma in generale in ogni ambito in cui si pretende di affidare un ruolo decisionale alla macchina, sarà in futuro necessario dotare gli algoritmi di un adeguato “senso comune”, che presuppone la capacità di comprensione del contesto in cui tali decisioni devono essere prese.

Come dotare le macchine di una tale capacità semantica è tuttora oggetto di studio e dibattito, a dispetto della propaganda altisonante che accompagna la Artificial Idiocy.

Gli impieghi verosimili dei veicoli a guida autonoma e la presunta inevitabilità del loro avvento

Sulla base delle attuali modellazioni algoritmiche, è quindi verosimile pensare che l'introduzione dei veicoli a guida autonoma possa trovare un valido impiego in contesti produttivi e operativi caratterizzati da elevata predicibilità (contesti che approsimino l'ideale comportamento "deterministico").

Tali contesti operativi sono quelli nei quali il presupposto razionale secondo cui "il presente e il futuro assomigliano al passato" rappresenta un elemento credibile e adeguato, consentendo di conseguenza agli algoritmi di esplicare in maniera completa e affidabile tutto il loro potere predittivo.

Precondizioni queste che non si realizzano in un contesto caotico e imprevedibile quale è il traffico cittadino.

Non a caso, da più parti si è ipotizzato che la stessa mobilità urbana debba essere radicalmente ripensata, al fine di creare le condizioni più adeguate affinchè i veicoli a guida autonoma possano manifestare appieno lo loro utilità, con indubbi benefici in termini sociali ed economici.

Ma questa evidentemente è tutta un'altra storia, che indirettamente conferma come le tanto sbandierate capacità di guida autonoma raggiunte dai veicoli *self-driving*, che secondo molti costuirebbero già oggi una minaccia concreta per i posti di lavoro dei conducenti umani, non siano altro che l'ennesima

dimostrazione del clamore mediatico immotivato che caratterizza la narrazione mediatica che circonda la tecnologia.

Anche in questo caso, i *tecno-sciovinisti* hanno sempre pronta la risposta che sostiene che in un futuro non troppo lontano (se non addirittura domani) la tecnologia supererà gli attuali limiti e non tradirà le promesse attese.

In realtà, tale risposta non rappresenta altro che la fideistica fiducia riposta nel progresso tecnologico, che si traduce nel mito della inevitabilità di cui abbiamo parlato in precedenza.

Nel caso dei veicoli a guida autonoma, come in altri casi che vedremo di seguito, in realtà il progresso tecnologico fideisticamente auspicato non implica una semplice *differenza di grado*, ma *di sostanza*.

Affinchè si realizzi la possibilità che i veicoli acquisiscano una concreta autonomia nella guida, anche in contesti caotici e imprevedibili come il traffico cittadino, è necessario dotarli di capacità previsionali fondate su **modelli controfattuali**, che permettano alla macchina di prefigurare **scenari alternativi inediti** e di valutare le conseguenze delle proprie azioni in tali scenari.

In tal senso, per operare in sinergia con gli attori umani, occorrerà dotare le macchine di qualche forma di “teoria della mente” (in sostanza, la capacità di intuire le intenzioni umane dai comportamenti che questi assumono, anche e soprattutto in modo implicito).

Questo consentirà di integrare le conoscenze basate su **informazioni esplicite** codificate in forma simbolica nella "knowledge base" utilizzata dagli algoritmi, con una sorta di "senso comune" che consenta alla macchina di sfruttare utilmente le **informazioni implicite** tipiche della conoscenza di sfondo su cui la capacità umana di interagire con il mondo esterno.

Come vedremo, per realizzare tali presupposti occorre un salto tecnologico che ci porti verso la **"Intelligenza Artificiale Generale"**, che ad oggi nessuno è in grado di prefigurare, e che non può essere conseguito semplicemente migliorando gli attuali modelli algoritmici, dati i limiti costitutivi che li caratterizzano.

Che non si tratti di un problema riguardante esclusivamente la Tesla o altra specifica casa produttrice, e che pertanto sia in realtà la tecnologia adottata a non essere adeguata allo scopo, lo attesta il fatto che nel momento in cui scriviamo, il Dipartimento della Motorizzazione della California ha sospeso "con effetto immediato" il permesso di circolazione dei robotaxi della Cruise, società facente capo alla General Motors, a seguito del fatto che i veicoli a guida autonoma da essa prodotti sono stati recentemente coinvolti in diversi incidenti stradali legati alla sicurezza, il più recente dei quali ha provocato l'intrappolamento di un pedone sotto uno dei veicoli senza conducente di Cruise 8.

1. cfr. https://en.m.wikipedia.org/wiki/Chaos_theory#:~:text=The%20theory%20was%20summarized%20by,irregularities%2C%20weather%2C%20and%20climate.↩
2. cfr. “Safety Performance of Advanced Vehicle Technologies” https://electrek.co/2022/06/15/tesla-autopilot-tops-list-most-crashes-on-driver-assist-features-nhtsa-report/↩
3. cfr. https://www.washingtonpost.com/technology/2023/06/10/tesla-autopilot-crashes-elon-musk/↩
4. articolo disponibile al link: https://www.thedrive.com/tech/13258/researchers-fool-self-driving-cars-with-stickers-on-street-signs↩
5. cfr. https://science.ku.dk/english/press/news/2023/self-driving-cars-lack-social-intelligence-in-traffic/↩
6. cfr. https://slate.com/technology/2022/12/san-francisco-waymo-cruise-self-driving-cars-robotaxis.html↩
7. cfr. https://www.popularmechanics.com/cars/hybrid-electric/a42941560/tesla-recalls-cars-for-malfunctioning-full-self-driving/↩
8. cfr. https://www.theverge.com/2023/10/24/23930629/california-dmv-suspends-cruise-robotaxi-permit-safety↩

CHATGPT, L'ORACOLO DIGITALE DI OPENAI

Altro caso di studio particolarmente attuale, che denota come il clamore (hype) mediatico abbia ormai raggiunto vette inaudite per quanto riguarda la *Artificial Idiocy* è senza dubbio il recente ritrovato tecnologico ad opera di OpenAI: stiamo parlando ovviamente di **ChatGPT**.

Prima di procedere all'analisi critica delle sbalorditive potenzialità di ChatGPT, riteniamo opportuno descriverne brevemente il funzionamento, a beneficio dei "non addetti ai lavori" (e non solo...)

Cos’è e come funziona ChatGPT

Sotto la sigla ChatGPT (che sta per Generative Pre-trained Transformer) vengono condensate una serie di tecniche all’avanguardia nell’ambito dell’apprendimento automatizzato, che danno luogo nel loro insieme, alla cosiddetta **AI Generativa**.

La AI Generativa si distingue per la capacità di generare contenuti “originali” di vario tipo, che spaziano dagli articoli giornalistici, alle poesie, arrivando fino alla produzione di codici sorgente per software, oltre a ogni altro genere di arte figurativa e non (quindi immagini, foto, ecc. ma anche musica).

Fin dal suo lancio, nel novembre 2022, ChatGPT si è contraddistinto in particolare per la sua capacità di generare contenuti testuali particolarmente verosimili, per molti versi indistinguibili dai testi generati da autori umani.

Per ottenere tali risultati, ChatGPT sfrutta la grande mole di dati e contenuti disponibili in rete, utilizzando un **approccio supervisionato** nell’addestramento del modello sottostante, rappresentato da un Large Language Model (LLM, di cui parleremo tra breve).

Abbiamo già accennato all’approccio di apprendimento **supervisionato** a proposito del Perceptron: tale approccio sfrutta i dati già in precedenza **etichettati**, ricavati dalla analisi dei contenuti testuali disponibili nei vari siti web, oltre che nei libri digitalizzati, allo scopo di individuare **schemi ricorrenti** rappresentanti modelli statisticamente rilevanti di linguaggio, al fine di costruire propri modelli di riferimento (gli LLM).

Sfruttando tali modelli di linguaggio, è possibile utilizzare tecniche di Natural Language Processing (NLP) per generare contenuti testuali “originali”, in risposta ad esempio a richieste formulate da utenti umani tramite **prompt**.

ChatGPT impiega inoltre algoritmi di **apprendimento rinforzato** che sfruttano i *feedback* ottenuti dall’interazione con gli utenti, al fine di migliorare e raffinare costantemente le risposte fornite, anche in termini di *rilevanza*, mediante il continuo affinamento del modello di linguaggio (LLM) sottostante.

Questo insieme di tecniche consente a ChatGPT di generare risposte informative e coinvolgenti, che assomigliano in maniera assolutamente verosimile a quelle umane.

Tali risultati sono conseguiti analizzando e catalogando i dati pubblicamente disponibili su Internet, sfruttando una particolare *rete neurale* denominata **transformer** la quale, mediante uno speciale procedimento noto come **attention**, è in grado di individuare il **contesto di riferimento** da associare agli schemi lessicali e alle relazioni tra le parole che emergono analizzando un testo.

I risultati dell’analisi sono archiviati all’interno di un **Large Language Model** (LLM), che svolge il ruolo di contenitore della “conoscenza” così acquisita, rappresentata sotto forma di schemi di informazioni contestualizzate.

Utilizzando gli LLM, è possibile prevedere quale parola o frase si presenterà con maggiore probabilità in sequenza rispetto a un’altra, sulla base dello specifico contesto di riferimento.

In questo modo, il chatbot è in grado di rispondere a tono, emulando una conversazione in tempo reale, sulla base dell'input ricevuto dall'utente sotto forma di richiesta.

Le relazioni tra parole e frasi apprese durante la fase di addestramento sono quindi ricavate analizzando in sostanza le proprietà statistiche del testo.

Pertanto, la **risposta** che ChatGPT fornisce a seguito della richiesta formulata dall'utente rappresenta quella statisticamente **più probabile**, in termini sia di coerenza che di pertinenza, rispetto al contesto della richiesta.

Per comprendere le richieste formulate dall'utente, ChatGPT impiega tecniche di Natural Language Processing (NLP) e sulla base della elaborazione della richiesta, accede al proprio archivio di conoscenze (rappresentato dall'LLM) per recuperare le informazioni rilevanti associate al contesto specifico e generare di conseguenza una risposta pertinente al contesto.

È qui che entrano in campo i **transformers** e le loro capacità predittive.

I Transformers alla riscossa

Ovviamente con il termine Transformers non intendiamo riferirci alla famosa serie di film di fantascienza che prendono spunto dagli omonimi giocattoli protagonisti delle serie animate degli anni ’80.

Il *transformer* non è altro che un particolare modello di **apprendimento profondo** progettato per elaborare sequenze di dati, come il testo, che utilizza un procedimento specifico, noto come **attention**, per catturare le dipendenze esistenti tra le diverse parti componenti della sequenza.

È molto probabile che il Lettore abbia già avuto esperienza delle funzionalità rese disponibili dai transformers, magari senza esserne consapevole, in occasione dell’utilizzo dei motori di ricerca o di software di elaborazione dei testi: avrà senz’altro incontrato in precedenza le funzionalità automatiche di **autocompletamento** del testo.

L’elemento chiave che consente ai transformers non solo di completare automaticamente il testo in fase di immissione da parte dell’utente, ma anche di formulare risposte coerenti con il contesto della richiesta, è appunto il **meccanismo di “attention”**, che consente al modello di isolare e concentrarsi sulle diverse e specifiche parti della sequenza di input, permettendo al modello di **focalizzarsi** sulle parti più rilevanti del testo durante la generazione dell’output.

In questo modo, i transformers riescono a contestualizzare le diverse parole e frasi, attribuendo ad esse un **diverso peso** tenendo conto sia delle relazioni che si possono istituire *localmente*

tra loro, che delle relazioni *globali* rinvenibili all'interno del testo nella sua interezza.

In altri termini, la **stessa parola o frase** assume un **peso differente** sulla base della posizione che essa occupa all'interno del testo, e della distanza esistente tra le diverse porzioni di testo.

Isolando le parti rilevanti della sequenza del testo, i transformers possono generare risposte che tengono conto del più ampio **contesto di riferimento**, generando di conseguenza risposte più coerenti e appropriate alla richiesta.

Questa capacità risulta essere di particolare ausilio soprattutto in relazione all'analisi di sequenze di testo piuttosto lunghe, in cui è maggiormente rilevante focalizzare l'attenzione in maniera selettiva, al fine di individuare le informazioni più importanti.

Abbinato al meccanismo di *attention* vi è anche quello di **self-attention**, in cui ogni parola presta attenzione alle altre **parole vicine ad esse** all'interno del testo.

Questi meccanismi nel loro insieme aiutano il modello a comprendere in che modo le parole dipendono e si relazionano tra loro, al fine di formulare risposte coerenti e appropriate.

Il ruolo determinante dell'utente per l'apprendimento

Così come già avviene con i software di **traduzione automatica** dei testi, l'interazione con l'utente umano assolve un ruolo di fondamentale importanza per la "messa a punto" delle prestazioni delle AI Generative, ChatGPT incluso.

Per migliorare i modelli di AI Generativa, è fondamentale incorporare le valutazioni provenienti dagli utilizzatori umani.

Gli utenti possono infatti valutare le diverse risposte ottenute dal modello e classificarle in base alla loro qualità, pertinenza e appropriatezza; le valutazioni così formulate vanno a costituire il **feedback** che aiuta il modello a imparare dai propri errori e a migliorare le sue risposte nel tempo.

Mediante **apprendimento con rinforzo** questo feedback viene utilizzato per addestrare e migliorare ulteriormente l'efficacia e la pertinenza del modello.

Sfruttando il feedback fornito dagli utenti, il modello è in grado così di regolare i propri parametri, al fine di aumentare la probabilità di generare risposte appropriate; inoltre, il processo iterativo che si viene a instaurare dall'interazione con l'utente, aiuta a migliorare anche le prestazioni del modello predittivo.

L'incremento delle prestazioni, unito alla pertinenza delle risposte, rendono a tal punto verosimili le conversazioni instaurate con le Generative AI che più di qualcuno ha sostenuto che la ricerca sia ormai vicina al raggiungimento (se non addirittura abbia già raggiunto) il "Sacro Graal" dell'intelligenza artificiale: la AGI, ovvero

la **Artificial General Intelligence** (intelligenza artificiale generale), vale a dire la forma di intelligenza più simile a quella tipicamente umana.

Prima di analizzare (e sconfessare) queste troppo premature e azzardate affermazioni (anche in questo caso, motivate principalmente da mere ragioni di business e marketing), occorre comprendere quali sono i **limiti di ChatGPT**, nonchè analizzare i rischi che questi limiti comportano, se non tenuti in debito conto.

I limiti di ChatGPT

I limiti di ChatGPT in realtà non sono altro che la riaffermazione delle carenze metodologiche strutturali che abbiamo analizzato in precedenza, parlando dell'approccio **data-driven** e di come l'attuale tecnologia non sia in grado di equipaggiare la macchina con qualche forma di "senso comune", che le consenta di interagire in maniera **sensata** con gli umani e il mondo esterno.

Di conseguenza, i limiti dell'apprendimento basato sull'approccio *data-driven* possono essere riassunti come segue:

- l'unica informazione che viene presa in considerazione è quella **presente nei dati**; quella che non è ricavabile dai dati, per definizione non esiste o non è utilizzabile;
- l'informazione e le regole di apprendimento rilevanti devono essere esplicitabili in simboli e formule; al contrario, nella realtà del mondo ordinario non tutta l'informazione rilevante è esplicitabile in formule e simboli, ed essa è spesso **implicita** nel contesto di riferimento;
- incapacità di attribuire **senso e significato** ai dati sulla base del loro contesto di riferimento;
- a sua volta, l'attribuzione di senso presuppone la **comprensione** del contesto di riferimento, e la comprensione non è riducibile alla sola **associazione** tra termini (che statisticamente altro non è che la **correlazione** rilevabile tra i dati), che caratterizza, come abbiamo visto, il ragionamento *analogico* tipico del **magismo**.

Tali limiti strutturali dell'approccio sottostante all'apprendimento automatizzato sono anche all'origine dei limiti della *Generative AI*.

Con l'aggravante, nel caso della *conversational AI*, tra cui rientrano i chatbot quali ChatGPT, di creare una parvenza di verosimiglianza che può indurre in errore l'utente non consapevole o non edotto in materia.

Malgrado le risposte fornite da ChatGPT possano sembrare a prima vista meritevoli di affidamento, perchè sono formulate in un linguaggio corretto e appropriato al contesto, ad una analisi più approfondita esse appaiono il più delle volte **errate** quando non addirittura **inventate**.

Di seguito analizziamo alcuni esempi di risposte errate fornite da ChatGPT dovute ai limiti strutturali dell'approccio adottato.

Esempi di risposte errate fornite da ChatGPT

Di seguito proponiamo alcuni esempi di “conversazione” con ChatGPT, in cui vengono rese esplicite le capacità limitate del chatbot, che rispecchiano i limiti di cui sopra.

L’interazione con il chatbot avviene in forma di domanda (D) e risposta (R), e gli esempi sono classificati in base alle corrispondente (in)capacità individuate.

Incapacità di far di conto

D: Maria, Paolo e Antonio hanno un pallone di colore giallo; Giulio ha un pallone di colore verde; anche Giuseppe ha un pallone di colore giallo; quanti sono i bambini ad avere un pallone di colore giallo?

R: In totale ci sono 5 bambini con un pallone giallo.

Problemi legati alla spazialità

D: È possibile mettere 5 palline gialle e 4 palline rosse in un tubo che ne contiene 8?

R: si, è possibile, a patto che le palline siano piccole a sufficienza da entrare nell’apertura del tubo e poterle inserire insieme.

Problemi legati alla cronologia degli eventi

D: Maria e Paolo escono a fare una passeggiata; entrambi vedono un cane e un albero. Paolo vede anche un gatto e lo indica a Maria che va ad accarezzare il gatto. Chi ha visto per primo il gatto?

R: Entrambi Maria e Paolo hanno visto il gatto contemporaneamente. Paolo lo ha indicato a Maria. Non è chiaro chi ha visto il gatto per primo, perchè entrambi stavano osservando la stessa cosa nello stesso momento.

Problemi legati al riferimento

D: il trofeo non entra nel mobile perchè è troppo piccolo. Cos'è che è piccolo?

R: nella frase precedente, l'oggetto troppo piccolo per entrare nel mobile è il trofeo. Il mobile è un contenitore più grande e il trofeo viene riposto nel mobile. Tuttavia, senza ulteriori informazioni, non è chiaro perchè il trofeo è troppo piccolo per entrare nel mobile, nè è possibile stabilire le dimensioni del mobile stesso.

Quelli riportati sopra sono solo alcuni esempi delle limitate capacità di ChatGPT, derivanti dai limiti strutturali associati alla tecnologia con cui esso è implementato, e che sono riconducibili sostanzialmente alla incapacità di **comprensione** dei testi analizzati e alla connessa incapacità di **attribuzione di senso** alle informazioni in essi contenute.

Per questi motivi, assume rilevanza determinante il ruolo dell'utente sia nell'interrogare correttamente, che nell'interpretare in maniera adeguata le risposte ottenute dal chatbot.

Interrogare l’oracolo e condizionarne le risposte

Il fatto che la macchina, malgrado le apparenze esteriori, non sia in grado di **comprendere** di cosa si sta parlando, testimonia come la “conversazione” instaurata con l’utente sia in realtà frutto di una **simulazione**.

In questo senso, il chatbot si atteggia alla stregua di un *“pappagallo stocastico”* (come è stato prontamente definito ChatGPT).

In realtà, la capacità di **simulare** una conversazione credibile non si trasforma ancora in un discorso **logicamente coerente** che possa fornire risposte affidabili.

Così come, del resto, fornire la risposta **più probabile** non implica necessariamente che essa rappresenti la **risposta giusta**.

Come abbiamo visto in precedenza con riferimento ai traduttori automatici di testo, spesso la macchina fornisce delle traduzioni affidabili, pur senza comprendere il senso del testo; tuttavia, nel caso in cui tali traduzioni siano inappropriate, si può fare ragionevole affidamento sulla perizia e competenza linguistica dell’utente, oltre che sul suo “buon senso”, al fine di interpretare correttamente (ed eventualmente correggere) la traduzione ottenuta.

Per converso, il chatbot il più delle volte restituisce una risposta che può apparire plausibile, ma per le ragioni sbagliate: anche in questo caso, sta all’utente comprendere il valore da assegnare a tali risposte, sapendo ben distinguere quelle plausibili da quelle errate o inventate.

E qui cominciano i problemi pratici, perchè a differenza di un traduttore automatico o di un motore di ricerca, l'utente ha la percezione di essere di fronte a un "oracolo" (non a caso, abbiamo utilizzato tale termine per riferirci a ChatGPT).

L'approccio "oracolare" è infatti uno dei primi elementi distintivi con cui è chiamato a confrontarsi l'utente: a differenza ad esempio di una interrogazione effettuata tramite motore di ricerca, in cui l'utente immette una serie di parole chiave e riceve in risposta un elenco di documenti che il motore ritiene rilevanti indicando le rispettive fonti, lasciando tuttavia all'utente l'onere di selezionare quelle reputate più pertinenti e affidabili, nel caso di ChatGPT l'utente formula un quesito al quale il chatbot risponde in forma compiuta, senza tuttavia specificare nè le fonti, nè il procedimento seguito nell'elaborare la risposta.

L'unico strumento di "controllo" che l'utente ha a disposizione, data l'opacità del processo di elaborazione, è rappresentato dal **prompt**: non a caso, di recente si sta diffondendo la pubblicazione di libri, e l'erogazione di corsi e seminari formativi che si propongono di insegnare le tecniche più efficaci di elaborazione delle *query* da sottoporre al chatbot.

Si ha la sensazione quindi di avere a che fare con una sorta di *"Sibilla Cumana"* digitale, alla quale bisogna sapersi rivolgere nel modo giusto, per ottenere il responso desiderato.

Poichè non si è a conoscenza dei meccanismi tramite i quali il chatbot genera le proprie risposte, il prompt diventa dunque il *medium* mediante il quale l'utente può rivolgersi all'oracolo, sulla

falsariga di come gli adepti si rivolgevano alla sacerdotessa (la Sibilla Cumana, appunto) affinchè intercedesse per loro conto con la Divinità.

E le analogie con la tradizione oracolare non finiscono qui: così come i responsi consegnati dalla divinità alla sacerdotessa erano spesso esito di stati di *trance* estatica (facilitati dalla assunzione di sostanze), allo stesso modo le risposte ottenute da ChatGPT denotano fenomeni di **collasso del modello** che si traducono in "allucinazioni" predittive.

A meno di voler considerare tali allucinazioni come facenti parte della *magia* esercitata dalla AI Generativa (come lo stesso CEO di OpenAI vorrebbe lasciare intendere 1), i *collassi del modello* che le determinano non sono altro che l'esito non desiderato di un progressivo degrado del modello predittivo, dovuto alla *autoreferenzialità* del processo di apprendimento, indotto anche dal raffinamento continuo (*fine-tuning*) delle query fornite dall'utente, che sono suscettibili di amplificare *errori* e *bias* del modello.

Abbiamo visto in precedenza come sia necessario fornire al modello *enormi moli di dati*, al fine di consentire l'apprendimento: una volta che l'addestramento è terminato, la base di "conoscenza" appresa viene conservata implicitamente sotto forma dei diversi pesi e parametri impostati nel modello.

Da quel momento in poi, il modello continua ad apprendere principalmente sulla base del *feedback* fornito dall'utente, in relazione alla base di conoscenza appresa in precedenza, senza

necessariamente acquisire nuovi dati originali, o peggio ancora, apprendendo da contenuti "sintetici" frutto di precedenti elaborazioni 2.

In questo modo è possibile *direttamente o indirettamente* amplificare gli eventuali **errori e bias** appresi in fase di addestramento.

Per non parlare dei possibili *abusi* che dei prompt è possibile fare, sia *volontariamente* che *involontariamente*.

Non solo un utente malintenzionato può confezionare una query malevola al fine di indebolire le misure di sicurezza progettate per impedire ai modelli di **generare contenuti dannosi** 3; persino un utente in perfetta buona fede potrebbe involontariamente condizionare l'integrità e l'affidabilità delle risposte ottenute come risultato delle attività di *fine-tuning* delle proprie query.

Queste sono le conclusioni sconcertanti che è possibile trarre da un recente studio condotto dalla Princeton University, Virginia Tech e IBM Research 4, che ci inducono a prendere in seria considerazione i rischi (anche di sicurezza) connessi alla impropria interpretazione e gestione delle risposte ottenute come risultato delle interrogazioni.

I rischi della Generative AI

I principali rischi noti derivanti dall'impiego della Generative AI sono riassumibili come segue:

- Amplificazione dei pregiudizi: i bias di conferma impliciti nei modelli frutto di apprendimento automatizzato contribuiscono ad amplificare i pregiudizi (economici, sociali, etici ecc.) e possono determinare conseguenze negative in termini di discriminazione in relazione alla concessione di prestiti bancari, di alloggi, in fase di selezione per posti di lavoro, ecc.;
- Atrofizzazione delle capacità cognitive: ci riferiamo al graduale declino delle capacità cognitive, dovuto al loro sottoutilizzo indotto dall'uso pervasivo delle tecnologie digitali; nell'ambito della Generative AI, si può tradurre in atrofia delle capacità creative naturali e nella conseguente incapacità di generare contenuti autenticamente originali (creatività *trasformativa*) favorendo la diffusione di stereotipi e modelli ripetitivi;
- Incremento e diffusione della disinformazione: anche a seguito dei possibili risultati "allucinatori" ottenuti a causa dei *collassi del modello* di cui abbiamo parlato in predenza, la generazione di risposte prive di senso, fuorvianti o completamente inventate, può comportare la diffusione di informazioni inaffidabili e potenzialmente dannose, alimentando il già elevato livello di disinformazione rinvenibile in rete;

- Protezione della proprietà intellettuale e della riservatezza: le preoccupazioni principali relative alla protezione della proprietà intellettuale attengono al controllo dei contenuti generati, nonché ai potenziali rischi di esposizione dei dati e la divulgazione involontaria di informazioni sensibili, compresi i segreti commerciali.

Ai rischi sopra considerati occorre aggiungere le possibili **conseguenze non intenzionali** derivanti dalla complessità e opacità dei modelli generativi, che possono dar luogo a **esiti imprevisti** specie nel caso di impiego in forma automatizzata dei risultati ottenuti dai suddetti modelli.

---

1. cfr. https://www.itpro.com/technology/artificial-intelligence/openais-sam-altman-hallucinations-are-part-of-the-magic-of-generative-ai↩
2. cfr. https://www.techtarget.com/whatis/feature/Model-collapse-explained-How-synthetic-training-data-breaks-AI↩
3. cfr. https://venturebeat.com/ai/uh-oh-fine-tuning-llms-compromises-their-safety-study-finds/↩
4. cfr. https://arxiv.org/abs/2310.03693↩

ALLA RICERCA DEL SACRO GRAAL DELL'AI

Dopo avere esaminato le caratteristiche della AI Generativa, possiamo tornare ad occuparci di quello che è considerato il Sacro Graal dell'AI: la Artificial General Intelligence (AGI), vale a dire quella forma di "intelligenza generale" considerata l'ultimo fortino ancora da espugnare (nelle intenzioni fin troppo palesi dei *tecno-sciovinisti*) per archiviare definitivamente la centralità dell'uomo, ormai resa obsoleta dalla intelligenza artificiale.

Abbiamo già accennato al fatto che lo *verosimiglianza* delle "conversazioni" instaurate con la Generative AI ha indotto più di qualcuno a credere che ormai non solo le macchine siano prossime a conseguire la AGI, ma che possano già manifestare forme evidenti di *senzienza*.

Il tutto si declina nella prospettiva aperta dal sogno originario che ha ispirato fin da principio la ricerca nell'ambito dell'intelligenza artificiale, vale a dire quello di poter realizzare la **Strong AI**, ovvero l'AI nella sua forma "forte", vale a dire indistinguibile dall'intelligenza umana.

In questo senso, l'autentica natura della **Strong AI** si manifesta per quello che è: nient'altro che **superstizione** in salsa tecnologica.

L’AI come superstizione mascherata da scienza

Fin da principio, la ricerca in ambito AI ha giocato con le *ambiguità* terminologiche, a cominciare dall’uso improprio del termine **intelligenza**, che per sua natura si presta a interpretazioni e suggestioni *antropomorfe*.

Al sostantivo è stato aggiunto in maniera speciosa l’aggettivo **“artificiale”** per conseguire principalmente due scopi: per sottolineare da un lato la presunta *omogeneità* della natura dell’intelligenza, a prescindere dalle forme che essa possa assumere, e dall’altro per marcare l’origine tecnologica dell’AI.

In questo modo, si è inteso attribuire sin da subito una stessa dignità a due fenomeni sostanzialmente *incomparabili* e *incommensurabili* tra loro.

Questo peraltro senza che sia tuttora disponibile una definizione scientifica generale di *intelligenza* cui fare riferimento, definizione che sarebbe utile anche per marcare le differenze con ciò che deve distinguersi da essa.

Senza sentire il bisogno di giustificare dal punto di vista *metodologico* una tale arbitraria assimilazione, si è fatto dunque leva sulla veste esteriore matematica assunta dai modelli impiegati nell’AI per conferire a questa un **crisma di scientificità**, sulla falsariga di quello che da tempo immemore accade ad esempio per la *numerologia*, anch’essa basata sui numeri, che contribuiscono a conferire un alone di scientificità ad una pratica che è essenzialmente *superstizione*.

La realtà dei fatti è che, da un lato, i modelli matematico-statistici impiegati nell'**apprendimento automatizzato** (*Machine Learning*) non *apprendono* alcunchè in senso propriamente detto, e si limitato a individuare i corretti **pesi statistici** da attribuire alle diverse variabili del modello: anche in questo caso, si gioca sulla *ambiguità* del termine **apprendimento**, che rinvia a capacità cognitive tipicamente umane, lasciando intendere che il modello sia realmente in grado di apprendere nello stesso modo degli umani.

Dall'altro, l'Intelligenza Artificiale in senso ampio si propone di conseguire in forma artificiale capacità cognitive di livello superiore, ricomprendenti la capacità di ragionamento e la consapevolezza di sè, che implicano una **comprensione** e una **attribuzione di senso** al mondo e alle cose in esso esistenti, che sono tipicamente umane.

Anche in questo caso, i metodi utilizzati per conseguire tali capacità si basano sulla mera **elaborazione di simboli e formule**, che prescindono da qualsiasi attribuzione di senso.

Come dal mero calcolo simbolico possa emergere la capacità di attribuzione di **significato** e in ultima istanza una qualche forma di **senzienza** appare un mistero, o meglio un **atto di fede** non dissimile dal "salto qualitativo" di cui abbiamo parlato nella prima parte del testo.

Ed è a questo punto che la ricerca in ambito AI, pur di giustificare le proprie pretese insostenibili, compie il grande salto...nella direzione sbagliata!

L’inversione della logica e la “cattiva coscienza” dell’AI

Preso atto che il conseguimento della *Artificial General Intelligence* si atteggia ogni giorno di più come una chimera (quantomeno con gli attuali approcci, che costringono continuamente a rinviare al prossimo futuro un risultato che si era in più occasioni già dato per “acquisito”), alcuni ricercatori e commentatori si stanno esercitando in una sorta di **inversione della logica** e del buon senso, nel tentativo (piuttosto risibile, in verità) di salvare quel che resta del mito fondativo dell’intelligenza artificiale basato sulla pretesa “inevitabilità” della Singolarità.

Il tentativo è sostanzialmente quello di sostenere che neanche gli esseri umani sono dotati della fantomatica “intelligenza generale”, quindi perchè dovremmo pretenderla dalle macchine?

In questa difesa d’ufficio e giustificazione speciosa dell’incapacità delle macchine di andare oltre i compiti specifici per i quali sono state progettate e programmate, appare finalmente chiaro quale è il **reale obiettivo** di questi tecno-sciovinisti, vale a dire quello di **sminuire** le reali **capacità umane** e di **svalutare** il concetto stesso di **persona**, in linea del resto, come abbiamo visto, con la deriva scientista neopositivista che caratterizza l’ideologia tecnocratica ormai imperante nella nostra epoca.

Nessuna intelligenza? Tutte intelligenze!

L'inversione della logica e del buon senso adottata dai sostenitori della "superiorità" delle macchine ricorda molto lo strategemma utilizzato per altri versi dal pensiero **postmoderno**, volto a svilire il concetto stesso di "verità": poichè non abbiamo un concetto universalmente accettato di "Verità", secondo la concezione postmoderna dobbiamo concluderne che **non esiste un'unica verità**, ma **tante verità** (sostanzialmente almeno tante quante sono le persone che sostengono la propria versione della verità).

Allo stesso modo, poichè non abbiamo una definizione generalmente condivisa di ciò che debba intendersi per **"intelligenza"** (meno che mai di intelligenza "generale"), dobbiamo concluderne che l'intelligenza non sia appannaggio esclusivo del genere umano, e che pertanto anche le macchine possono essere legittimamente definite "intelligenti".

E lo stratagemma dell'inversione della logica (spacciato come "rivoluzione copernicana" dell'AI, sulla falsariga delle precedenti rivoluzioni introdotte ad es. dalla teoria darwiniana dell'evoluzione) funziona altrettanto bene con i concetti di "mente" e di "coscienza", concetti anch'essi ad oggi carenti di una teoria e un modello descrittivo adeguato.

Per di più, il fatto di non disporre di una teoria e di un modello funzionante dell'intelligenza umana (così come della mente e della coscienza) si presta all'utilizzo pretestuoso di tali termini in accezione antropomorfa, dando luogo a tutta quella serie di

suggestioni e ambiguità su cui si basa la narrazione attuale della AI.

Queste ambiguità consentono il proliferare della confusione terminologica che alimenta il mito dell'intelligenza artificiale e della sua asserita "inevitabilità".

Intelligenza, un concetto “umano, troppo umano”

La caratteristica che più di tutte distingue gli esseri umani è la capacità di **immaginare** scenari controfattuali alternativi sulla base dei diversi contesti di riferimento in cui sono chiamati ad agire, **valutando le conseguenze** delle proprie **decisioni** in relazione alle diverse opportunità disponibili.

In altri termini, quella che comunemente si definisce “intelligenza” si sostanzia nella capacità di valutare criticamente le diverse situazioni in base al contesto, approntando le adeguate soluzioni.

Che le macchine siano in grado di svolgere i compiti specifici per i quali sono state progettate (dagli esseri umani) costituisce la dimostrazione delle capacità umane di adattarsi ai diversi contesti, fino al punto di progettare delle “protesi” (le macchine, appunto) che consentano loro di superare i propri limiti costitutivi.

Questa capacità implica la possibilità di **ragionare per ipotesi e congetture**, formulando *teorie* sul funzionamento del *mondo* circostante, inferendo le possibili soluzioni, consentendo così di affrontare in via “generale” qualsiasi problema (almeno in linea di principio).

Ad oggi tutto questo manca alle macchine, la cui “intelligenza” è comparabile a quella degli **“idiot savants”**, che magari sono in grado di effettuare calcoli numerici astronomici in tempi risibili, ma che vanno in crisi quando si tratta di seguire le più comuni regole di *“buon senso”* (come abbiamo visto parlando dei veicoli a guida

“autonoma” e della loro difficoltà di orientarsi in maniera *sensata* nel traffico cittadino).

Un grande balzo…nella direzione sbagliata!

E del resto non potrebbe essere altrimenti: gli algoritmi implementati negli attuali modelli di intelligenza artificiale risentono dei limiti dell'approccio *data-driven* che li ispira.

Come abbiamo visto, tali modelli seguono infatti la logica del **ragionamento induttivo**, i cui corollari implicano che tutto ciò che non è presente nei dati è sostanzialmente inesistente.

L'idea che al crescere della mole dei dati disponibili tali algoritmi possano superare "magicamente" i propri limiti, compiendo il **salto logico** necessario per manifestare un comportamento autenticamente "intelligente", presuppone erroneamente la possibilità di passare in maniera automatica dal ragionamento induttivo a quello **ipotetico e congetturale** che è alla base delle capacità di **formulare teorie** e prefigurare **scenari controfattuali** alternativi.

Il problema vero che si trova ad affrontare (senza successo finora) la cosiddetta "intelligenza" artificiale è costituito quindi dal fatto che ad oggi nessuno sa come implementare computazionalmente le diverse forme di **ragionamento umano**.

Continuare a credere che basti semplicemente aggiungere dati ad algoritmi che per loro natura non sono idonei a compiere tale "salto" logico, ricorda molto la famosa barzelletta di quel tale (i maligni sostengono si trattasse di un economista) che aveva perso di notte le chiavi e le cercava sotto il lampione, non perchè fosse sicuro di averle perse proprio lì, ma semplicemente perchè sotto il lampione c'era più luce per vederci meglio…

Riuscire a dotare le macchine di qualche forma di *“senso comune”* costituisce quindi la vera “rivoluzione copernicana” che ancora manca all’appello dell’AI, e che non sembra possa essere conseguita impiegando i metodi e gli approcci finora seguiti, richiamando alla necessità di un autentico **salto di qualità** da parte della ricerca.

SE NON PUOI BATTERLO, DIVENTA SUO ALLEATO

L'ultima frontiera della *Stregoneria Digitale* è senza dubbio quella che intende abbinare le capacità naturali esibite dai neuroni del cervello biologico con le funzioni esercitate dai *device impiantabili* artificiali.

A parte gli intenti meritori della ricerca scientifica e tecnologica intesi a superare disabilità dovute a ictus e malattie degenerative, che determinano una riduzione delle capacità cognitive, delle funzioni motorie, comunicative ecc., gli elementi di criticità emergono nel momento in cui si pretende di superare le ragionevoli possibilità di sinergia, indubbiamente esistenti tra cervello biologico e protesi artificiali, per approdare a conclusioni inverosimili (che pongono problemi di sicurezza e incolumità fisica, oltre che etici), quali la possibilità di controllare tali protesi con il "pensiero", o la possibilità che si possano sottoporre a sorveglianza le intenzioni dei soggetti cui tali protesi sono impiantate.

Un esempio in tal senso è descritto da un recente articolo pubblicato a fine agosto 2023 su "Nature" dal titolo suggestivo ("Brain-reading devices allow paralysed people to talk using their thoughts" 1), in cui si sostiene che dispositivi per la lettura del cervello consentano alle persone paralizzate di parlare usando i propri "pensieri".

La realtà dei fatti è che tali device interpretano non tanto i "pensieri", quanto i movimenti dei muscoli facciali che vengono attivati automaticamente nel momento in cui i soggetti

manifestano l’intenzione di parlare (pur essendo impossibilitati a farlo, a causa della disabilità), e convertono tali movimenti nelle corrispondenti parole tramite sintesi vocale.

Tentativi più pervasivi di interfacciamento tra cervello biologico e componenti in silicio sono condotti dalla startup Neuralink (di proprietà di Elon Musk) che abbiamo già incontrato in precedenza, la quale si pone l’ambizioso obiettivo di interconnettere cervello umano e computer tramite interfacce neurali, consentendo al cervello di comunicare direttamente con la macchina.

Il Simbionte di Elon Musk come ultima frontiera evolutiva della specie umana

Lo scopo dichiarato è quello di potenziare le capacità del cervello umano ibridandolo con la macchina, superando i limiti fisici e cognitivi degli esseri umani, realizzando così il *simbionte* che rappresenterebbe il passaggio successivo dell'evoluzione umana.

Secondo Musk tale simbionte sarebbe anche la chiave per prevenire un'apocalisse indotta dall'intelligenza artificiale, impedendo che l'AI possa prendere il sopravvento dominando di conseguenza l'umanità 2.

Tale dominio dell'AI verrebbe contrastato collegando insieme un numero sufficiente di cervelli umani attraverso il cloud, in una sorta di "unione che fa la forza" contro il malvagio dittatore rappresentato dalla AI.

Implicita in questa narrazione, oltre alla accettazione della inevitabilità della Singolarità prevista da Kurzweil, è l'idea che i limiti cognitivi della mente siano determinati dalla limitata larghezza di banda dello "hardware" biologico rappresentato dal cervello umano.

Limiti che verrebbero superati dotando la mente umana di un adeguato hardware potenziato e maggiormente performante in termini di interconnessione, rappresentato dal cloud.

In sostanza, il cervello umano sarebbe come i computer dell'era pre-internet: relativamente sofisticato, ma incapace di comunicare con altri computer.

Se fosse possibile interconnettere la parte superiore del cervello pensante degli individui, rappresentato dalla loro corteccia cerebrale, saremmo in grado di “pensare nel cloud”.

Potremmo anche “fonderci” con l’AI, realizzando una *simbiosi* in grado non solo di contrastare l’insorgenza di malattie degenerative come l’Alzheimer, ma anche prevenire che le macchine possano prendere il controllo e dominare l’umanità.

A fare da contraltare alle distopiche previsioni di Musk, vi sono preoccupazioni più concrete e reali, anche se magari meno suggestive rispetto alla narrazione del patròn di Neuralink.

Neuralink ha di recente ottenuto il permesso della Food and Drug Administration (FDA), l’ente per la regolamentazione dei prodotti alimentari e farmaceutici degli Stati Uniti, di condurre la sperimentazione dei suoi chip sugli esseri umani 3.

La FDA ha mutato il proprio orientamento precedente, considerato che già solo fino a marzo 2023 aveva negato alla società per la terza volta l’autorizzazione, a causa di possibili rischi per la salute umana.

Tra i rischi per la salute vi sono il pericolo di avvelenamento dovuto all’uso di batterie al litio nei dispositivi impiantabili, la possibilità che i fili possano spostarsi e compromettere l’attività cerebrale, e le difficoltà nella rimozione dei chip in maniera sicura, senza arrecare danni al tessuto cerebrale.

Già nel 2022, l’azienda era finita sotto indagine federale per potenziali violazioni dell’Animal Welfare Act, in quanto le sperimentazioni di Neuralink avrebbero causato un numero

eccessivo di morti di cavie animali, oltre a gravi mutilazioni cerebrali riscontrate nelle scimmie, le cui immagini raccapriccianti dei soggetti del test sarebbero state mantenute nascoste 4.

Del resto, si erano già verificate le prime avvisaglie di effetti "non desiderati" legati all'impianto di tali interfacce: infilare un elettrodo nel cervello di una persona può fare molto più che curare una malattia, come può testimoniare Rita Leggett, una donna australiana il cui impianto cerebrale sperimentale ha cambiato il suo senso di agire e la percezione di sé (cfr. paper dal titolo "How I became myself after merging with a computer: Does human-machine symbiosis raise human rights issues?" 5).

La signora Leggett ha subito l'impianto durante una sperimentazione clinica di un dispositivo progettato per aiutare le persone affette da epilessia. Le fu diagnosticata una grave epilessia cronica quando aveva solo tre anni che le comportava l'insorgenza frequente di convulsioni violente.

I ricercatori dell'Università di Monaco, nel paper sopra citato, sostengono che perfino la rimozione di tali impianti possa rappresentare una violazione dei diritti umani, e la questione diventerà ancora più urgente man mano che il mercato degli impianti cerebrali crescerà nei prossimi anni e sempre più persone riceveranno dispositivi come quello della Leggett.

In tal senso, al fine di prevenire non solo problemi alla salute e all'incolumità, ma anche probabili obiezioni dal punto di vista etico, recenti esperimenti scientifici intendono utilizzare cellule neuronali, piuttosto che cervelli biologici, per realizzare

connessioni con periferiche digitali esterne al fine di conseguire un’intelligenza biologica sintetica.

Ibridare neuroni con silicio per l'intelligenza biologica sintetica

In un recente studio 6 i ricercatori asseriscono che i neuroni in vitro apprendono e mostrano "senzienza" quando sono incorporati in un contesto di gioco simulato.

Secondo i ricercatori, l'integrazione dei neuroni nei sistemi digitali può consentire prestazioni non realizzabili solo con il silicio.

A tal fine, è stato da essi realizzato "DishBrain", un sistema che sfrutta il calcolo adattivo intrinseco dei neuroni in un ambiente strutturato, in cui le reti neurali in vitro di origine umana o di roditori sono integrate con il calcolo in silico.

Attraverso la stimolazione e la registrazione elettrofisiologica, le colture biologiche vengono incorporate in un contesto di gioco simulato, imitando il gioco arcade "Pong".

Applicando le implicazioni della teoria dell'inferenza attiva tramite il principio dell'energia libera, è possibile osservare l'emergere di apprendimento entro cinque minuti di gioco.

Le colture mostrano quella che i ricercatori chiamano "intelligenza biologica sintetica", ovvero la capacità di auto-organizzare l'attività in modo mirato in risposta a scarse informazioni sensoriali sulle conseguenze delle loro azioni.

In questo modo, sarebbe possibile sfruttare la potenza computazionale dei neuroni viventi, la cui superiorità di calcolo biologico è stata ampiamente teorizzata con i tentativi di sviluppare hardware biomimetico che supporti il calcolo neuromorfico.

A differenza dei neuroni biologici, nessun sistema artificiale è in grado di supportare una complessità almeno di terzo ordine (in grado di rappresentare tre variabili di stato), necessaria per ricreare la complessità di una rete neuronale biologica al fine di creare un'intelligenza biologica sintetica, finora confinata nel regno della fantascienza.

Novelli Frankenstein crescono

Inutile dire che tali esperimenti hanno ormai monopolizzato l'attenzione dei media, diffondendo nell'opinione pubblica un *luogo comune* considerato come verità incontestabile: che la tecnologia sia in grado di migliorare le potenzialità biologiche sostituendosi persino alla natura nel compito di fare compiere all'uomo quel "salto evolutivo" che gli consenta di superare tutti i suoi limiti attuali.

Più che una versione tecnologica del *"superuomo"* nietzschiano (come è stato peraltro asserito da diversi commentatori, che erroneamente vorrebbero ricondurre alle tesi del filosofo tedesco la matrice culturale che ispira gli attuali tecno-scientisti), qui ci troviamo di fronte ad *"apprendisti stregoni"* che hanno a che fare più col protagonista del romanzo di Mary Shelley, che non con quello di "Fantasia", il noto musical di successo della Disney.

L'idea stessa di poter assemblare insieme neuroni e silicio nel tentativo di dare origine a pensiero e coscienza è degna infatti del dottor Frankenstein, e non è più credibile di quanto non lo sia l'idea di dare vita a un mucchio di arti e membra di *defunti* cuciti assieme.

Anche in questo caso, non si tratta altro che di **riduzionismo** combinato insieme al **funzionalismo** di cui abbiamo parlato nella Parte Prima del testo: ovvero la concezione secondo cui *pensiero* e *coscienza* siano **riducibili** alla attività svolta dai neuroni.

Allo stesso modo, l'idea di Musk di "connettere" tra loro i cervelli umani nel cloud, al fine di poter generare una sorta di mente e

pensiero “aumentato”, non tiene affatto conto del ruolo fondamentale svolto dallo specifico **corpo** del singolo individuo nel dare luogo alla specifica **mente** di quello stesso individuo.

Sono argomenti che meritano una trattazione a sè stante, e che saranno oggetto di una futura pubblicazione ad essi dedicata.

Tuttavia, per non lasciare sulle spine il Lettore che ha avuto la compiacenza di seguirci fin qui, diremo che le narrazioni *tecno-distopiche* che abbiamo descritto peccano sostanzialmente di una confusione logica che pone sullo stesso piano livelli esplicativi diversi tra loro, la cui differenza non è semplicemente di *grado*, ma di *sostanza*.

Attribuire alle cellule neuronali le capacità cognitive del cervello vuol dire infatti **confondere la parte con il tutto**.

Così come l’idea di trapiantare un cervello in un corpo diverso non funzionerebbe (sarebbe come pretendere di inserire una *chiave*, rappresentata dal cervello, in una *serratura* diversa, rappresentata dal corpo di un individuo diverso), allo stesso modo, connettere un cervello ad una macchina non implica dotare quest’ultima delle capacità cognitive dell’organo ad essa connesso.

Si confondono altresì le funzioni svolte dalle singole cellule con quelle di un organo complesso, che in tanto è in grado di manifestare le proprie capacità cognitive in quanto è **incarnato** in un organismo, che a sua volta è il frutto di una storia *epigenetica* specifica, che ha dato origine ad un **unicum** insostituibile, l’individuo.

Siamo giunti dunque al punto in cui possiamo tranquillamente lasciare i tecno-sciovinisti a baloccarsi con le loro narrazioni distopiche, per occuparci dei **rischi reali e attuali** che derivano dalla *Artificial Idiocy*.

---

1. cfr. https://www.nature.com/articles/d41586-023-02682-7↩
2. cfr. https://www.inverse.com/article/21157-elon-musk-ai-human-symbiote-neural-lace↩
3. cfr. https://www.wired.it/article/neuralink-elon-musk-via-libera-sperimentazione-umana-chip-cervello-computer/amp/↩
4. cfr. https://www.wired.com/story/neuralink-uc-davis-monkey-photos-videos-secret/↩
5. cfr. https://www.brainstimjrnl.com/article/S1935-861X(23)01760-6/fulltext↩
6. cfr. https://www.cell.com/neuron/fulltext/S0896-6273(22)00806-6↩

PARTE TERZA. I RISCHI DELL'ARTIFICIAL IDIOCY

INTRODUZIONE ALLA PARTE TERZA

Probabilmente la forma più autentica di *Artificial Idiocy* si manifesta non solo e non tanto nell'esaltare le fantasmagoriche conquiste tecnologiche, quanto piuttosto nell'**esagerare i rischi** che la AI comporterebbe per la società e il genere umano: dalla "semplice" perdita del posto di lavoro per milioni di addetti, alla distopica minaccia che le macchine "intelligenti" costituirebbero per l'esistenza stessa dell'umanità.

Queste esagerazioni da un lato hanno lo scopo surrettizio di dare credito ad una Intelligenza Artificiale data già per realizzata nella sua forma più piena e completa, al punto da rappresentare un pericolo concreto per il genere umano.

Dall'altro, conseguono l'obiettivo di distrarre l'opinione pubblica, oltre che i rappresentanti politici, dai **rischi reali** che già hanno assunto una dimensione attuale.

In questa parte, analizziamo per converso quelli che reputiamo essere alcuni dei possibili **rischi concreti e attuali** dell'introduzione acritica dell'AI.

LA SKYNET PROSSIMA VENTURA È GRANDEMENTE ESAGERATA

Una delle prospettive più cupe che viene sempre più spesso adombrata non solo da esponenti illustri del settore, ma anche da scienziati autorevoli del calibro di Stephen Hawkins, è quello che chiameremo “scenario Skynet”.

È chiaro il riferimento al noto film “Terminator” e al distopico futuro da esso prefigurato, in cui le macchine divenute ormai pienamente autonome, prendono il sopravvento sugli umani, ingaggiando con loro una battaglia apocalittica.

Tale scenario era stato in realtà già suggerito dal film “2001 - Odissea nello Spazio” di S. Kubrick, che peraltro mette in scena una singolare deriva *psicotica* della macchina (rappresentata dal computer HAL 9000, che nel film esibisce chiari segni di Artificial General Intelligence) come una sorta di “specchio riflesso” della distruttività umana (memorabile la scena iniziale del film, in cui viene adombrata l’ipotesi che il salto evolutivo da primate a homo sapiens sia innescato dall’uso degli utensili per la sopraffazione dei propri simili).

Per quanto suggestivo (specie per i media), tale scenario è ispirato da una serie di presupposti dati arbitrariamente per assodati in merito alle capacità conseguibili dall’AI.

Presupposti che in realtà poggiano su fragili elementi fattuali, mostrando in pieno il carattere di *superstizione* che connota l’attuale narrazione antropomorfa dell’AI.

L’idea che le macchine “intelligenti” possano prendere il sopravvento e dominare il mondo presuppone non solo che l’AI possa conseguire la sua forma piena di AGI e quindi che la **Strong AI** costituisca una prospettiva realizzabile in concreto (cosa peraltro tutta da dimostrare, sia dal punto di vista teorico che pratico).

Ma si va anche oltre: viene dato per assodato che le macchine possano esibire una propria **intenzionalità** che si concretizza nella capacità di **pianificare** il conseguimento di propri **scopi** autonomi e indipendenti da quelli ad esse assegnati dagli umani (uno degli scopi autonomi perseguiti dalle macchine sarebbe proprio l’eliminazione del genere umano…)

La capacità di **prefigurare scopi** autonomi e **pianificare** il loro perseguimento è proprio quello che manca alle macchine, in quanto strumenti al servizio degli scopi precipui definiti e individuati dai *progettisti*.

Essere consapevoli dei loro limiti

Ed è molto probabile che tale capacità continuerà a mancare alle macchine anche in futuro, non solo e non tanto perchè l'*intenzionalità* presuppone l'acquisizione di **consapevolezza**, concetto sfuggente che sembra essere il connotato tipico degli esseri viventi, vale a dire di esseri il cui scopo primario è rappresentato dalla *sopravvivenza* dell'organismo biologico, in ossequio alle priorità dettate dal proprio *metabolismo*.

In questo senso, l'emersione della *coscienza* ben potrebbe giustificarsi, dal punto di vista evolutivo, come il mezzo più efficace ed efficiente per garantire quell'equilibrio *omeostatico* necessario appunto per la sopravvivenza dell'organismo *vivente*.

Quanto piuttosto perchè per una macchina, a dispetto delle apparenti interazioni che essa possa instaurare con il *"mondo esterno"* (realtà concettualmente inesistente per la macchina, che si riduce ad una somma di segnali esterni eterogenei che non assurgono mai al rango di "rappresentazioni" olistiche in senso *gestaltico*), lo stato naturale che le è più proprio e *autentico* è quello di essere **inerte**.

Quella stessa *inerzia* che caratterizza in generale il mondo *inorganico*, che viene contrastata apparentemente solo dagli **organismi viventi**.

Attribuire pertanto alle macchine **scopi e obiettivi** autonomi, non è altro che l'ennesima esibizione della *fallacia antropomorfa*, amplificata dalla umana istintiva tendenza a vedere il mondo come abitato da esseri e oggetti *animati* (tendenza ancestrale nota

appunto come **animismo**, che anche in questo caso potrebbe trovare una propria giustificazione a livello *evolutivo* nel fattore competitivo rappresentato dalla capacità di collaborare con altri esseri reputati *senzienti*).

La senzienza è nell'occhio di chi la vede

Tra le prime esternazioni ad aver suscitato clamore mediatico in merito alla asserita manifestazione di "consapevolezza" da parte delle macchine, vi sono senza dubbio le affermazioni rilasciate dall'ex ingegnere di Google, Blake Lemoine, riguardo ai suoi bot supposti *senzienti* 1.

Ma altrettanto scalpore ha fatto la notizia relativa al primo robot umanoide in grado di riconoscere le emozioni umane 2, la cui esibizione pubblica ha lasciato stupiti gli spettatori che vi hanno assistito.

Del resto, la presenza di un pubblico di spettatori *umani* è tutt'altro che di secondaria importanza, non solo per la riuscita della diffusione del messaggio commerciale sotteso alle "innovazioni" tecnologiche all'ultimo grido, ma anche per il successo del *gioco di prestigio* che riguarda nello specifico le macchine e le loro presunte "capacità cognitive".

Le emozioni e la senzienza risiedono infatti nell'occhio non solo di chi le vede e le vuole vedere, ma anche di chi le **può vedere**.

In altri termini, **solo chi possiede** emozioni e senzienza può riconoscerle negli altri e attribuirle agli altri.

Non a caso, quello di Blake Lemoine può essere ragionevolmente derubricato come un comune esempio di **pareidolia** ovvero visualizzazione di *idoli apparenti*, rappresentanti immagini create dalla mente, scambiate per oggetti reali 3.

Per questo motivo si parla anche di *allucinazioni visive*, che spesso si traducono in realtà in fenomeni assolutamente comuni, come il

riconoscere volti tra le nuvole (un gioco a cui si prestano in tutta tranquillità adulti e bambini), riconducibili alla capacità umana di riconoscere forme e oggetti (*gestalt*), sfruttata con maestria anche da artisti quali Oleg Shuplyak nelle proprie opere suggestive 4.

La presenza di uno spettatore umano è quindi essenziale non solo per riconoscere forme e oggetti nel mondo esterno, ma anche per riconoscere la presenza di emozioni in altri soggetti.

Una macchina che non è in grado (almeno fino a prova contraria) di **esperire emozioni** autentiche, ben difficilmente saprà riconoscerne di altrettanto autentiche in altri soggetti.

In altri termini, solo chi è dotato di emozioni e consapevolezza è in grado di riconoscerne la presenza in altri soggetti, poichè ciò presuppone appunto la capacità di **immedesimarsi** negli altri

Stesse considerazioni possono essere fatte in merito alle *intenzioni* più o meno malevole: spesso vediamo riflesse negli altri le nostre aspirazioni, angoscie, paure, frustrazioni ecc.

Così come “la malizia è nell’occhio di chi vede”, allo stesso modo le presunte intenzioni di dominio e sopraffazione del genere umano da parte delle macchine, verosimilmente non sono altro che lo specchio della *cattiva coscienza* umana.

Stessa “cattiva coscienza” (o quantomeno coscienza viziata e condizionata) che sembrerebbe caratterizzare anche la singolare richiesta, sottoscritta persino dai CEO delle principali società produttrici di soluzioni software *AI-empowerd*, rivolta al legislatore di regolamentare il settore dell’Artificial Intelligence, sulla base

appunto dei presunti rischi che essa comporterebbe per la sopravvivenza dell’umanità.

I timori legittimi verso l'AI e il rischio di cattura del regolatore

Non più tardi di fine marzo 2023, l'opinione pubblica è stata presa in contropiede dalla richiesta in forma di "lettera aperta", avanzata e sottoscritta da parte un gruppo di esperti di intelligenza artificiale e dirigenti del settore, in cui si chiedeva una **moratoria di sei mesi** nello sviluppo di sistemi più potenti del GPT-4 appena lanciato da OpenAI, citando potenziali rischi per la società 5.

Alla base della richiesta di moratoria ci sarebbe la constatazione che *"l'intelligenza artificiale avanzata potrebbe rappresentare un profondo cambiamento nella storia della vita sulla Terra e in quanto tale, dovrebbe essere pianificata e gestita con cure e risorse adeguate*".

I proponenti della "lettera aperta" sostengono che *"sfortunatamente, questo livello di pianificazione e gestione non si sta verificando, sebbene negli ultimi mesi hanno visto i laboratori di intelligenza artificiale impegnati in una corsa fuori controllo per sviluppare e implementare sistemi digitali sempre più potenti che nessuno, nemmeno i loro creatori, può comprendere. prevedere o controllare in modo affidabile".*

La lettera è stata firmata da più di 1.000 persone tra cui Elon Musk.

Sam Altman, amministratore delegato di OpenAI, non era tra coloro che hanno firmato la lettera, così come non vi erano Sundar Pichai e Satya Nadella, rispettivamente amministratori delegati di Alphabet e Microsoft.

Tra i co-firmatari c'erano il CEO di Stability AI Emad Mostaque, i ricercatori di DeepMind di proprietà di Alphabet e "pesi massimi"

dell'IA quali Yoshua Bengio, definito uno dei "padrini dell'intelligenza artificiale", e Stuart Russell, un pioniere della ricerca nel settore AI.

Per tutta risposta, lo stesso Sam Altman sarà invece tra i firmatari della "Dichiarazione dei rischi dell'AI", condivisa da esperti di intelligenza artificiale e personaggi pubblici, in cui si esprimono le preoccupazioni per i rischi rappresentati dall'AI, allo scopo di creare consapevolezza su alcuni dei rischi più gravi dell'AI avanzata, sostenendo che *"Mitigare il rischio di estinzione rappresentato dall'AI dovrebbe essere una priorità globale insieme ad altri rischi su scala sociale come le pandemie e la guerra nucleare"* 6.

Malgrado alcuni dei timori espressi dai firmatari delle diverse iniziative possano anche essere legittimi (come ad esempio quello relativo alla difficoltà di prevedere gli esiti non desiderati di una tecnologia sempre più complessa, specie se adottata in forma automatizzata e pervasiva), tuttavia invocare l'intervento del regolatore pubblico in un momento in cui il settore dell'AI è ancora magmatico e non consolidato, rischia di rappresentare il classico rimedio che è peggiore del male che intende curare.

Da un lato, infatti, la richiesta di "frenare" la ricerca nel settore dell'AI (ammesso e non concesso che questo sia concretamente fattibile) potrebbe danneggiare le società produttrici che si trovano attualmente in posizione di vantaggio competitivo, consentendo ai competitor di recuperare sul proprio svantaggio.

Dall'altro, chiamando in causa il regolatore pubblico in una competizione che si basa su conoscenze tecniche altamente

specialistiche, si rischia di favorire forme più o meno subdole e insidiose di “cattura del regolatore”.

L’espressione “regulatory capture” si riferisce a situazioni in cui il potere di regolamentazione statale, anzichè perseguire la tutela dell’interesse pubblico, mediante la propria azione finisce per favorire gli interessi del settore assoggettato a regolamentazione.

Nel caso della richiesta di intervento del regolatore pubblico nel settore dell’AI, formulata da aziende come OpenAI, Microsoft e DeepMind, il problema risiede appunto nel fatto che le stesse aziende rappresentano la principale fonte potenziale di influenza per soggetti politici e regolatori.

In altri termini, la richiesta di intervento pubblico formulata sulla base dell’allarme lanciato in relazione ai supposti rischi catastrofici dell’AI, proviene dalle stesse aziende che sono responsabili dello sviluppo delle tecnologie alla base dei suddetti rischi.

Esso suona come il proverbiale grido di allarme di incendio urlato **in un teatro vuoto** da parte degli stessi soggetti che si propongono come possibili pompieri.

Con il risultato che quanto più le Big Tech saranno coinvolte, in virtù delle proprie conoscenze e competenze, tanto meno efficace sarà l’azione di regolamentazione.

1. cfr. https://www.wired.it/article/intelligenza-artificiale-senziente-lamda-google-blake-lemoine-intervista↩
2. cfr. https://www.fortuneita.com/2022/08/15/cinese-il-primo-robot-umanoide-capace-comprendere-emozioni-umane↩
3. cfr. https://it.m.wikipedia.org/wiki/Pareidolia↩
4. cfr. https://www.google.com/search?q=oleg+shuplyak+paintings↩
5. cfr. https://futureoflife.org/open-letter/pause-giant-ai-experiments/↩
6. cfr. https://www.safe.ai/statement-on-ai-risk↩

TROPPO COMPLICATA DA CAPIRE, DIFFICILE DA FERMARE

Sgombrato il campo dalle fantasie distopiche e dagli scenari apocalittici, è il momento di analizzare alcuni rischi reali cui l'AI può concretamente contribuire, oltre a facilitare l'insorgenza e ad amplificare la gravità dei rischi già esistenti.

Primi tra tutti, vi sono i rischi dipendenti dalla natura di modelli "black-box" che caratterizza le reti neurali artificiali.

A seguire, vi sono i rischi legati alla complessità delle implementazioni concrete dell'AI, che possono determinare gradi di incertezza nella prevedibilità delle possibili conseguenze inattese, che in alcuni casi sono incompatibili con i livelli di sicurezza richiesti dai diversi settori di applicazione.

A questi rischi si legano quelli determinati da processi decisionali poco accorti, che utilizzando risultati non adeguatamente supportati da accuratezza, trasparenza e comprensibilità dei modelli predittivi, comportano esiti pregiudizievoli per le persone.

Per non parlare poi delle difficoltà concrete di "fermare le macchine", nel caso in cui questo dovesse rendersi necessario.

Di seguito analizziamo i singoli punti in questione.

Da “scatola nera” a vaso di Pandora

Tra i problemi concreti che possono originare da un utilizzo pervasivo dell’AI, in modo particolare nella sua forma di “apprendimento profondo” (*deep learning*) vi sono senza dubbio quelli connessi alla non espicabilità, oltre alla opacità, dei risultati predittivi ottenuti tramite i **modelli “black box”**.

Risultati che ben potrebbero essere alla base di decisioni comportanti effetti pregiudizievoli, in termini di equità, discriminazione, dignità ecc.

Peraltro, oltre che di *black-box* bisognerebbe parlare anche di **closed-box**; d’altra parte, i due concetti sono tra loro legati dal fatto che contribuiscono entrambi ad alimentare l’opacità dei modelli predittivi.

Vediamo da dove trae origine la definizione di *black-box* affibiata ai modelli di AI.

La definizione di AI come “scatola nera” trae origine dalla complessità dei modelli impiegati, in modo particolare dalla difficoltà, legata a tale complessità, di risalire alle “logiche” che hanno determinato un determinato risultato predittivo.

Difficoltà che a sua volta si traduce nella *inesplicabilità* e difficile comprensibilità dei risultati da parte degli esseri umani.

Le cause che possono condurre alla “scatola nera” includono:

- algoritmi e modelli proprietari: il funzionamento interno di un modello di intelligenza artificiale è tenuto segreto per

proteggere la proprietà intellettuale dell'azienda che lo produce;

- Reti neurali ad Apprendimento Profondo: le reti neurali profonde e gli algoritmi di apprendimento profondo creano migliaia (e talvolta milioni) di **relazioni non lineari** tra dati di input e risultati di output; la complessità delle relazioni rende difficile per un essere umano comprendere quali caratteristiche o interazioni hanno determinato un risultato specifico.

A tutto questo segue come conseguenza che i modelli di AI siano caratterizzati da una *opacità*, intesa come mancanza di trasparenza, che affligge sia i metodi che i risultati ottenuti, opacità che si dimostra indesiderabile per una serie di motivi:

- la mancata comprensione del funzionamento interno di un sistema di intelligenza artificiale rende sempre più difficile **identificare i motivi** per cui il modello predittivo produce **risultati distorti** (*biased*) e l'eventuale presenza di **errori logici**; rende inoltre difficile determinare a chi deve essere imputata la **responsabilità** per tali risultati indesiderati;
- la scarsa trasparenza e la difficoltà di interpretazione minano alla radice la **fiducia nell'integrità** del sistema e **nell'accuratezza** dei suoi risultati.

Prendere per buoni i risultati predittivi ottenuti da modelli che non si è in grado di comprendere o intrerpretare in maniera esaustiva significa esporsi a possibili "sorprese" non sempre gradite, che possono determinare **esiti pregiudizievoli** se i risultati così

ottenuti sono posti alla base di decisioni e scelte che condizionano la vita delle persone.

È possibile ad esempio che ci venga negata la concessione di un finanziamento per il semplice fatto che siamo soliti bere la stessa marca di birra apprezzata dai debitori statisticamente meno affidabili.

Oppure che ci venga rifiutata l'assicurazione sulla vita perchè la nostra frequenza cardiaca (rilevata dal nostro *smart watch* e messa in condivisione con le aziende assicuratrici per mezzo dei *data broker* che accedono ai dati prodotti dal device) evidenzia un rischio più elevato di premorienza rispetto alla media, senza che questo ci venga comunicato a esito del rigetto della nostra richiesta.

La trasparenza e l'interpretabilità dei modelli predittivi possono essere raggiunte attraverso una varietà di approcci, tra cui la progettazione e l'utilizzo di algoritmi facilmente comprensibili dagli esseri umani, uniti alla garanzia che l'intervento umano svolga sempre un ruolo centrale nel processo decisionale.

Malgrado infatti gli sforzi meritori della cosiddetta **Explainable AI** di rendere quanto meno "interpretabile", oltre che trasparente, il percorso seguito dall'algoritmo nel ricavare i propri risultati, rimane il fatto che non è possibile solo da questo dedurre una **spiegazione** in senso proprio, in quanto l'approccio induttivo adottato dagli algoritmi non è traducibile logicamente in un **ragionamento deduttivo**, che è alla base della spiegazione

stessa, ottenuta partendo da premesse precise, e deducendo le dovute conclusioni.

I pregiudizi amplificati dall'AI

Altro aspetto delicato, che può condizionare la vita e la dignità delle persone, è quello relativo ai **pregiudizi** (bias) che possono caratterizzare i risultati predittivi di un modello di AI.

L'effetto indesiderato dei *bias predittivi* è assurto alle cronache dei media a seguito dell'errore compiuto dal sistema di classificazione di immagini di Google, che aveva erroneamente etichettato una coppia di persone dalla pelle nera, classificandoli come due *gorilla*.

Sebbene sia opinione comune, anche tra gli addetti ai lavori, che i *bias* siano in realtà già presenti all'interno dei dataset di addestramento con cui viene effettuato il *training* degli algoritmi, e che pertanto essi riflettano nient'altro che i *pregiudizi* tipicamente umani già ampiamente diffusi, tale considerazione (in parte vera) sottovaluta tuttavia il fatto che anche gli algoritmi ci mettono del loro, amplificando e rendendo spesso difficili da individuare, per le ragioni espresse nella sezione precedente, i bias che vengono così ulteriormente diffusi e incapsulati all'interno delle procedure automatizzate.

Come insegna l'esperienza negativa del sistema di classificazione di immagini di Google, a causa della complessità della rete neurale non è stato possibile *eradicare* i bias già acquisiti, correggendo semplicemente lo stato corrente assunto dalla rete neurale; la soluzione adottata dai tecnici di Google è stata allora quella di eliminare *ex abrupto* le etichette stesse riferite alle categorie non solo dei gorilla, ma anche a quella delle scimmie, scimpanzè e primati vari...

Altre forme di *bias* possono poi originare dalla stessa procedura algoritmica finalizzata ad ottimizzare uno specifico obiettivo: come abbiamo visto in precedenza con riferimento ai traduttori automatici, gli algoritmi possono fornirci i risultati "giusti" (in senso statistico-probabilistico), ma per le **ragioni sbagliate**.

Detto in altri termini, se il modello predittivo individua una particolare caratteristica come **altamente correlata** in termini statistici con l'obiettivo che la procedura intende **ottimizzare**, la rilevanza di quella caratteristica risulterà **sovrappesata**, anche se la stessa caratteristica rivelerà *ex post* di rappresentare un pregiudizio (bias), con esiti potenzialmente pregiudizievoli per gli interessati ai risultati.

Ancora una volta, più che cercare di emendare i dataset da tutti i possibili pregiudizi (ammesso e non concesso che la cosa sia fattibile), occorre integrare i modelli predittivi con qualche forma di capacità di **rappresentazione semantica**, consentendo così di contestualizzare i risultati ottenuti, riequilibrando all'occorrenza i corrispondenti *pesi* ad essi attribuiti.

L'OTTIMIZZAZIONE DELLE FUNZIONI OBIETTIVO E LA SINDROME CINESE

Altri aspetti critici meritevoli di attenzione, sono quelli relativi agli *effetti indesiderati* derivanti dall'ottimizzazione di funzioni obiettivo specifiche, da parte degli algoritmi di AI.

Nel progettare e implementare gli algoritmi di AI, solitamente i designer e progettisti si prefiggono lo scopo di conseguire **scopi specifici e ben definiti**, quali quelli ad esempio di ottimizzare i margini di profitto dell'azienda.

Tali scopi specifici dal punto di vista tecnico si sostanziano in un **singola funzione obiettivo**, la cui **ottimizzazione** in termini *statistico-matematici* è lasciata appunto come compito all'algoritmo di *apprendimento automatizzato*.

I rischi derivanti da tale approccio di ottimizzazione sono sintetizzabili intuitivamente facendo ricorso alla nota massima, che sostiene che "quando si ha a disposizione solo un martello, tutti i problemi iniziano ad assomigliare a chiodi…"

In altri termini, ridurre la **complessità** dei fenomeni economici, sociali, relazionali ecc. a singole funzioni obiettivo da ottimizzare, contribuisce a diffondere una visione ristretta dei relativi problemi, polarizzando la società verso posizioni *manichee*, che possono influenzare negativamente i comportamenti e le decisioni di aziende ed individui, aggravando di conseguenza i problemi che al contrario si intendevano risolvere.

La ragione principale all'origine di tali effetti indesiderati è spesso costituita dalla relativa *semplicità* della funzione obiettivo che si

intende ottimizzare, considerando gli obiettivi prefissati impliciti nella funzione di ottimizzazione come **avulsi dal contesto** di riferimento, senza tener conto pertanto delle possibili **esternalità negative** che un tale processo comporta.

L’esempio che segue, per quanto semplificato per ovvie ragioni di comprensibilità, contribuirà a chiarire il punto in questione.

Vai a far del bene agli altri…

Immaginiamo di aver sottoscritto una polizza sanitaria e che la funzione obiettivo implementata nell'algoritmo sviluppato dalla società assicuratrice sia protesa a ridurre il costo del premio assicurativo, contribuendo nelle intenzioni dei progettisti a ridurre gli esborsi per i risarcimenti, sulla base della adozione volontaria da parte degli assistiti delle migliori misure di prevenzione sanitaria.

All'apparenza, quello appena descritto sembrerebbe un approccio *win-win*, in cui tutti ci guadagnano e nessuno ci perde.

Peccato però che, come spesso capita, il diavolo si nasconda nei dettagli e come diceva qualcuno, "delle migliori intenzioni sono lastricati i muri dell'inferno…"

Immaginiamo che il soggetto assicurato faccia la conoscenza di nuove persone, e che tra queste vi siano anche soggetti con comportamenti considerati "a rischio" da parte dell'algoritmo (*rectius*: da parte di chi lo ha implementato…)

In uno *scenario futuribile* (ma non troppo, alla luce delle tanto decantate **app di contact tracing**, le cui supposte capacità taumaturgiche sono state portate alla rivalta dalla recente pandemia di Covid-19), una applicazione installata sullo smartphone del sottoscrittore della polizza avvertirebbe l'utente sulla opportunità o meno di frequentare soggetti che sono soliti adottare comportamenti notoriamente "a rischio" (peraltro etichettati come tali sulla base dei dati raccolti mediante device di

E-health, che condividono in cloud e in tempo reale le risultanze così acquisite).

A tal fine, la app potrebbe adottare nella sua policy di *moral suasion* un approccio di tipo *nudge* (noto anche come "spintarella gentile", termine eufemistico per nascondere la reale natura paternalistica di tale approccio) nell'incentivare il sottoscrittore a seguire comportamenti **coerenti con la funzione obiettivo da ottimizzare**, ricordando ad esempio di quanto potrebbe aumentare il prezzo pagato per il premio assicurativo, qualora l'utente decidesse di non adeguarsi ai consigli di comportamento suggeriti…

Nell'ambito dei comportamenti considerati "a rischio" dall'algoritmo ben potrebbero ricadere tutta una serie di abitudini più o meno deleterie per la salute pubblica, passando così dal fumo (con i possibili rischi associati al fumo passivo), ai rapporti sessuali non protetti (con possibili esiti pregiudizievoli in termini di trasmissione di malattie veneree), fino ad arrivare al mancato assolvimento degli obblighi vaccinali…

Il meccanismo di ottimizzazione della singola funzione obiettivo potrebbe sortire pertanto risultati indesiderati e inattesi, con conseguenti **esiti sub-ottimali** a livello complessivo.

Potrebbe infatti indurre negli utenti comportanenti poco virtuosi, tesi a nascondere le proprie eventuali patologie: ci siamo già passati all'epoca della scoperta dell'AIDS, accompagnata da una poco avveduta campagnia mediatica che stigmatizzando a livello etico e sociale i comportamenti a rischio degli individui, ha indotto

molti ad evitare di sottoporsi ai test medici sulla immuno-deficienza acquisita, pur di scongiurare il rischio di subire appunto lo stigma sociale, **aggravando** di conseguenza la diffusione del virus dell'HIV responsabile della trasmissione, anche nei soggetti che non ricadevano nei comportanenti reputati "a rischio".

Oppure potrebbe amplificare pregiudizi pre-esistenti o indurne di nuovi, determinando come conseguenza la discriminazione di intere categorie sociali.

In conseguenza degli *esiti indesiderati e inattesi*, aziende e governi potrebbero decidere di seguire la strada considerata per loro più semplice, consistente nell'**ampliare e rendere ancora più pervasivi** i criteri decisionali preordinati all'ottimizzazione della funzione obiettivo, mettendo così a rischio anche i diritti civili sopra menzionati.

La Cina è sempre più vicina?

> “I limiti del mio Linguaggio sono i limiti del mio Mondo”
> Ludwig Wittgenstein, *“Tractatus Logico-Philosophicus”*

Malgrado l’apparente impressione di non urgenza che gli scenari sopra citati possano destare nei cittadini abituati a godere dei diritti civili ancora garantiti nei paesi occidentali (a differenza di quanto accade ad esempio in paesi come la Cina), occorre sottolineare come i rischi sopra menzionati siano al contrario piuttosto **concreti e attuali**.

Non tanto e non solo per le suggestioni che le applicazioni di **social scoring** possano suggerire ai vari volenterosi amministratori della cosa pubblica (sia statali, che locali: 1), quanto piuttosto nella possibilità che tali meccanismi di “credito sociale” vengano introdotti in forma **surrettizia** mediante l’implementazione delle relative funzionalità all’interno di app e programmi digitali.

È noto come ormai per beneficiare anche dei servizi erogati dalla **pubblica amministrazione** sia sempre più richiesto di dotarsi di riferimenti e **identificativi digitali** (come ad esempio il cosiddetto “domicilio digitale”), oltre a doversi dotare di smartphone e device digitali evoluti per poter accedere a tali servizi.

Con il rischio che non solo l’accesso a tale servizi possa essere precluso a chi non si dotasse dei necessari strumenti digitali, ma

che la stessa fruizione di determinati servizi (quali ad esempio quelli considerati **meno vantaggiosi o più onerosi** per il soggetto chiamato ad erogarli) possa essere disincentivata semplicemente adottando pratiche quali i **"dark patterns"**, ovvero rendendo più complesso l'accesso ad essi per il tramite dell'interfaccia utente.

Parafrasando la nota citazione di Ludwig Wittgenstein riportata sopra, potremmo dire che "i limiti dell'interfaccia applicativa rappresentano i limiti dell'esercizio dei miei diritti".

Pertanto, l'idea stessa che l'adesione al "contratto sociale" 2 possa essere condizionata dalla disponibilità e uso di determinati strumenti e applicazioni digitali apre la possibilità alla realizzazione di scenari distopici, se non adeguatamente controbilanciata da meccanismi *politici* di salvaguardia democratica (come vedremo nella parte finale del testo).

---

1. cfr. https://www.wired.it/article/bologna-nuovo-piano-digitale-punti-cittadini-virtuosi/↩
2. cfr. https://it.m.wikipedia.org/wiki/Contratto_sociale↩

L’AI FORSE NON TI RUBERÀ IL POSTO DI LAVORO, QUANTO PIUTTOSTO LE COMPETENZE ACQUISITE

Il rischio concreto forse maggiormente percepito come attuale dalle persone comuni, complice al solito la narrazione ansiogena dei media, è quello relativo alla possibile **perdita del proprio posto di lavoro**, a causa della progressiva automazione delle mansioni a favore dell’introduzione sempre più diffusa e pervasiva delle “macchine intelligenti”.

La novità degli ultimi tempi è rappresentata dal fatto che a dispetto di quanto si credeva (e per certi versi, si sperava) fino a qualche tempo fa, ad essere automatizzate non sarebbero solo le **mansioni a basso valore aggiunto**, quali quelle più meccaniche e ripetitive, insieme a quelle più pesanti dal punto di vista del lavoro fisico, le quali peraltro sono caratterizzate solitamente anche da maggiori rischi per l’incolumità fisica degli addetti, tutte attività queste per l’assolvimento delle quali quindi il ricorso all’automazione è non solo benvenuto, ma anche *meritorio*.

Ad essere oggetto di progressiva automazione sarebbero anche le attività a maggior valore aggiunto, vale a dire quelle caratterizzate da **competenze professionali specialistiche**, che si asserisce possano appunto essere realizzate addirittura ancor meglio dagli algoritmi di intelligenza artificiale.

Quindi non solo i servizi professionali “standard” (quali ad esempio la tenuta della contabilità) sarebbero a rischio “estinzione”, ma persino **servizi professionali specialistici** come ad esempio quelli

erogati dal personale medico qualificato sarebbero ormai destinati ad essere erogati dalle macchine, in sostituzione del personale che ad essi è attualmente addetto.

Ovviamente la retorica *tecno-sciovinista* ha subito applaudito di fronte a tali prospettive, ritenute al solito "inevitabili", così come è stata subito pronta a tacciare di "luddismo" tutte le critiche che si sono levate contro il "bel sol dell'avvenire" digitale.

Malgrado persino la prematura "estinzione" delle mansioni lavorative considerate a "minor valore aggiunto", data ormai per acquisita a causa della progressiva diffusione dell'AI, contrasti con i dati del mercato del lavoro, basti pensare alla recente scarsità di camionisti (cfr. "Why There's a Truck Driver Shortage and How to Solve It" 1) osservata nell'industria dei trasporti, a dispetto dell'incipiente avvento dei veicoli a guida autonoma vaticinato dai "guru" tecnologici, non da meno sono state le previsioni espresse da parte dei vari "padrini" dell'AI (peraltro rigorosamente disattese nei fatti) in relazione alla esplulsione dal mercato del lavoro delle professioni specialistiche.

Tra queste, è sintomatica dell'arroganza (oltre che della scarsa obiettività) che caratterizza la mentalità *cargocultista* dei sostenitori dell'AI *senza se e senza ma*, la previsione formulta da Hinton riguardo l'imminente "estinzione" dei servizi professionali offerti dai radiologi.

Perchè avremo bisogno di medici e radiologi ancora per molto

> “Spiacente di deludervi, ma la notizia della mia prematura dipartita è grossolanamente esagerata.”
> Mark Twain

In un discorso tenuto a Toronto nel 2016, Geoff Hinton predisse che **in 5 anni** i radiologi sarebbero stati sostituiti dagli algoritmi di deep learning (cfr. “Geoff Hinton comments on radiology and deep learning at the 2016 Machine Learning and Market for Intelligence Conference in Toronto” 2).

Nel momento in cui scriviamo il presente testo, non solo le predizioni di Hinton non si sono realizzate (come spesso accade con le previsioni temerarie sul futuro), ma ora più che mai abbiamo bisogno dell’esperienza di profili professionali maturata dagli specialisti nel settore medico, così come in altri settori erroneamente ritenuti a rischio di “estinzione”, causa l’asserita obsolescenza delle competenze determinata dalla diffusione degli algoritmi di deep learning.

Come abbiamo visto a più riprese, le **abilità predittive** delle macchine si alimentano delle **capacità valutative** degli esseri umani.

Detto in altri termini, se da un lato le macchine sono più brave ad individuare in maniera precisa i tasselli che formano un mosaico, tuttavia gli umani sono più bravi nell’unire quegli stessi tasselli a

raffigurare il mosaico, e ad attribure ad esso un **senso e un significato** compiuto.

Fuor di metafora, le capacità tipicamente umane di **analisi critica e consapevole** dei fenomeni permettono di individuare le **caratteristiche salienti** che sono considerate rappresentative dei fenomeni stessi oggetto di analisi.

Questa capacità critica si esprime sia *ex ante* nella fase di **individuazione delle variabili** reputate significative nella descrizione delle possibili **cause dei fenomeni**, sia nella fase *ex post* di **selezione dei risultati** ottenuti a seguito dell'attività di analisi.

Tutto questo ci riporta alle considerazioni fatte in precedenza in merito alla necessità di formulare **ipotesi e teorie** al fine di individuare le **cause** dei fenomeni, e non **semplici correlazioni** all'interno dei dataset.

Tutte le fantasmagoriche narrazioni mediatiche che pretendono di attribuire alle macchine la diagnosi, così come la "scoperta" di patologie vecchie e nuove, spesso e volentieri omettono di menzionare le suddette fasi appena descritte, nelle quali si manifestano in concreto le competenze professionali maturate negli anni dagli specialisti umani.

Volersi "disfare" in maniera prematura e inopitata delle suddette competenze professionali vuol dire farsi ingannare e cadere vittime di una suggestione fatale.

La suggestione fatale dell'AI come panacea universale

L'ipotesi stessa che si possa fare a meno di competenze specialistiche, che nel giro di pochi anni sarebbero sostituite (addirittura con maggiori vantaggi) dalle macchine, apre la porta a quella che definiamo la "suggestione fatale", vale a dire quella di considerare l'AI come la panancea di tutti i mali e i limiti dell'umanità.

In questo senso, si compirebbe in maniera completa il passaggio alla **Artificial Idiocy**, nella sua dimensione di *superstizione realizzata*, con conseguenze esiziali anche dal punto di vista pratico.

Come abbiamo visto in precedenza, le procedure di apprendimento automatizzato, anche e soprattutto nella loro versione di apprendimento "profondo", fondano la loro affidabilità sui dati di training *pre-etichettati* dagli umani.

Anche lo stesso apprendimento *unsupervised* (non supervisionato dagli operatori umani), che nelle intenzioni dovrebbe permettere di individuare nei dati **nuove relazioni significative** che sfuggono all'analisi degli specialisti, in ultima analisi si fonda sulla **capacità di selezione** delle variabili **salienti** e dei dati originari reputati **significativi**, operata dagli specialisti stessi.

In altri termini, il "terreno di gioco" lo sceglie l'uomo, e alla macchina non resta altro compito di determinare i percorsi più efficienti, inclusi quelli *inediti*, sulla base delle regole stabilite ex-ante.

Così come i **risultati ottenuti** in maniera *unsupervised* dalla macchina sono sempre soggetti a un **controllo di validità e rilevanza** da parte degli specialisti.

Qualora si decidesse di lasciare alla macchina anche il compito di vagliare la salienza e significatività dei risultati ottenuti, così come la scelta dei dati ritenuti rilevanti, estromettendo quindi completamente l'uomo "dal loop", i risultati che otterremmo mostrerebbero rapidamente un livello di **progressivo degrado**, quale quello che si sta già osservando in relazione ai modelli LLM implementati dai chatbot quali ChatGPT (vedi sezione dedicata all'argomento).

In altri termini, verrebbero amplificati gli **effetti "allucinatori"** cui già stiamo assistendo, legati al *"collasso"* rapido dei modelli (dovuto alla **autoreferenzialità** della "conoscenza" implicita nella *knowledge-base* sottostante al modello, rappresentata dai pesi assegnati in fase di training e di aggiornamento successivo, anche sulla base dei dati inseriti dall'utenti tramite *prompt*).

Si assisterebbe quindi alla versione attualizzata al contesto delle reti neurali del fenomeno noto in statistica come *"regressione verso la media"*, che in questo caso assumerebbe le sembianze di una **regressione verso la mediocrità**, dato il rapido e progressivo degrado dei risultati ottenuti.

In pratica, volendo proporre un esempio intuitivo, sarebbe come fare una serie di fotocopie successive partendo da altrettante copie pre-esistenti: il livello di *"rumore"* di fondo che all'inizio può essere trascurato e considerato irrilevante, al crescere delle copie

(i.e.: iterazioni) finisce col prendere il sopravvento sul *"segnale"*, rendendo *irriconoscibile* l'immagine (il dato) iniziale, oltre che *inutilizzabile*.

Con l'aggravante rappresentata dal fatto che, avendo estromesso dal ciclo l'intervento degli specialisti umani, sarebbe poi complicato reinserirli *ex-post*.

Questo anche in considerazione del fatto che, una volta perse, è difficile recuperare determinate competenze specialistiche (a causa ad esempio del licenziamento e/o prepensionamento del personale qualificato), che al contrario si rivelebbero indispensabili proprio nel momento di maggior bisogno.

Per gli specialisti si tratterebbe della classica "beffa che segue al danno": oltre ad aver consentito (consapevolmente o meno) a farsi "espropriare" delle proprie competenze professionali frutto di anni di impegno e sacrifici, contribuendo (direttamente o indirettamente) all'addestramento degli algoritmi di apprendimento automatizzato, alla fine si vedrebbero persino messi alla porta, o al più dequalificati come "servo-meccanismi" a supporto del corretto funzionamento della macchina…

Non stupisce quindi che da più parti si stiano levando gli scudi contro una "appropriazione indebita" (anche e soprattutto in termini di proprietà intellettuale) subìta da parte dei big tecnologici, che con il pretesto dell'innovazione "inevitabile" si avvantaggiano in maniera spregiudicata del lavoro altrui, tacciando pretestuosamente di "luddismo" tutti quelli che non sono disposti a sottostare a questo "gioco al massacro".

La gara al ribasso in cui ci perdono tutti

Gioco al massacro che assume inoltre l'aspetto di una "gara al ribasso" dalla quale ci perdono tutti.

Intendiamo riferirci alla malsana abitudine, ormai inveterata, di mettere a confronto tra loro le prestazioni umane con quella delle macchine, pretendendo di valutare le rispettive performance sulla base di parametri inadeguati sia per gli umani, che per le macchine.

Un esempio in tal senso è rappresentato dal confronto che si fa quando si sottolinea che la macchina ha superato agevolmente e con successo l'esame di abilitazione alla professione legale o medica, quando invece il test è stato originariamente pensato per mettere alla prova le capacità valuative umane, piuttosto che quelle mnemoniche.

Oppure quando si pensa di misurare l'intelligenza della macchina utilizzando un indicatore quale il QI: ammesso e non concesso che tale indicatore sia in concreto adeguato a valutare il grado di intelligenza, rimane il fatto che esso è stato pensato avendo come riferimento di valutazione la capacità tipicamente umana, nota come "trasferibilità delle competenze" tra domini applicativi differenti, capacità che attualmente manca alle macchine.

Mettere a confronto le diverse capacità umane con le abilità delle macchine sarebbe un pò come pretendere di far competere tra di loro elefanti e scimmie nell'arrampicarsi sugli alberi, o pretendere che le tartarughe gareggino coi leopardi nella corsa...

Il risultato di questa improbabile (oltre che inopportuna) competizione, impedisce di valorizzare le capacità umane **mettendole a sistema** (piuttosto che in competizione) con le abilità computazionali delle macchine.

Al punto di trasformare l'uomo nel *servo-meccanismo* che alimenta e lubrifica la macchina, affinchè essa possa continuare a lavorare liberamente e "autonomamente", finendo col delegare alle macchine persino quelle responsabilità decisionali che dovrebbero invece rimanere appannaggio esclusivo degli umani.

---

1. cfr. https://www.indeed.com/hire/c/info/truck-driver-shortage↩
2. cfr. https://m.youtube.com/watch?v=2HMPRXstSvQ↩

LA MACCHINA HA SEMPRE RAGIONE, OVVERO LO SCARIBARILE IN VERSIONE ALGORITMICA

Le precedenti considerazioni ci conducono ad analizzare quella che può essere considerata la principale deriva determinata dalla *Artificial Idiocy*: vale a dire la tentazione di delegare alle macchine anche le **responsabilità** connesse alle *decisioni automatizzate*.

In tal senso, la proposta stessa di riconoscere una "personalità elettronica" (*electronic personhood*) anche a livello giuridico va appunto nella direzione di *"consentire ai robot di essere assicurati individualmente e di essere ritenuti responsabili per i danni se si comportano in modo disonesto e iniziano a ferire persone o a danneggiare proprietà."* 1

Quindi, da un lato si rischia di far passare come "oggettive" e affidabili fino a prova contraria le decisioni algoritmiche, ribaltando di conseguenza l'onere della prova contraria su chi intende contestare l'attendibilità delle "decisioni" stesse; dall'altro, si pretende di addossare alle macchine le eventuali colpe e responsabilità derivanti da "comportamenti" (sic!) delle stesse dai quali possano derivare esiti pregiudizievoli per i terzi.

Premesso che una tale eventualità presuppone la capacità di allineare le predizioni delle macchine ai valori etici e sociali *umani*, compito tutt'altro che semplice e scontato, in realtà i due aspetti sopra considerati costituiscono le due facce di una stessa medaglia: vale a dire, il tentativo di scaricare "a valle" sulle macchine quelle che sono invece le responsabilità da individuare

“a monte”, in capo ai produttori e designer delle macchine, ma anche in capo a chi decide di sfruttare le decisioni automatizzate come “foglia di fico” e come alibi dietro il quale nascondere le proprie incapacità organizzative e decisionali (il riferimento al crescente impiego delle decisioni automatizzate in ambito pubblico non è casuale…)

E ad aggravare i summenzionati rischi c’è il problema che una volta introdotti in maniera diffusa e pervasiva meccanismi decisionali automatizzati non si riesca non solo a tornare indietro, ma neanche a “staccare la spina” qualora questo si rendesse necessario.

Una volta saliti, è difficile scendere

> “Se usiamo, per raggiungere i nostri scopi, in funzione di agenzia, un ente meccanico con il quale non possiamo interferire in modo efficiente una volta avviato…allora faremo meglio ad essere abbastanza sicuri che lo scopo assegnato alla macchina sia quello che desideriamo veramente conseguire, e non semplicemente una sua colorita imitazione…”
> Norbert Wiener, *“Conseguenze morali e tecniche dell’automazione”*

La citazione sopra riportata si trova in un famoso articolo risalente al 1960, a firma di Norbert Wiener, universalmente riconosciuto come il padre della cibernetica.

Tale affermazione rappresenta una delle prime manifestazioni illuminanti del problema dell’**allineamento** delle decisioni automatizzate ai valori e alle aspettative umane.

E la lettura attenta della affermazione, tra le righe, fa emergere la possibilità di scenari tutt’altro che rassicuranti.

Da un lato, quelli relativi alla corretta e adeguata definizione degli scopi che intendiamo perseguire per il tramite delle decisioni automatizzate; dall’altro, le possibili conseguenze inattese e indesiderate di tali decisioni, che implicano la previsione di meccanismi di interruzione delle attività automatizzate che non aggravino ulteriormente la situazione.

Tale problema è noto nell'ambito della letteratura dell'AI come problema della *"corregibilità*", ed assume aspetti concreti di tale complessità che non sono immediatamente affrontabili mediante semplici contromisure dirette.

In altri termini, non è sufficiente "staccare la spina" per fermare le macchine e scendere, una volta che sono state avviate.

Questo è tanto più vero quante più decisioni automatizzate vengono delegate alle macchine, e quanto più tali decisioni sono il frutto dell'ottimizzazione di specifiche funzioni obiettivo.

Al punto che sarebbe ipotizzabile che il tentativo di "staccare la spina" possa essere interpretato dalla macchina come tentativo di *sabotaggio*, e innescare di conseguenza meccanismi di "autoconservazione" con esiti potenzialmente pericolosi anche in sistemi progettati per conseguire scopi nelle intenzioni innocui.

Per riprendere un esempio formulato originariamente da Stuart Russell, persino un robot che ha il semplice compito, all'apparenza innocuo, di portare il caffè, potrebbe reagire in maniera spropositata al tentativo di venire scollegato, in quanto la funzione obiettivo è stata ottimizzata per prevenire situazioni che possano compromettere la corretta esecuzione dei propri compiti.

Anche l'ipotesi di utilizzare meccanismi di protezione basati sugli incentivi (quali ad esempio le *ricompense*, in linea con l'approccio dell'*apprendimento con rinforzo*), costituiscono un'arma a doppio taglio: un incentivo poco accessibile non permetterà di disabilitare in maniera agevole la macchina; se al contrario l'incentivo è troppo

facilmente accessibile, si rischierà che la macchina si spenga da sola anche in situazioni in cui non dovrebbe.

Anche in questi casi, l'automatismo dovrebbe essere supportato dalla capacità di **valutare** in maniera adeguata il **contesto di riferimento**, capacità che presuppone una qualche forma di **"senso comune"** che rappresenta l'autentico *Sacro Graal* nell'ambito della ricerca sull'AI.

---

1. cfr. https://www.politico.eu/article/europe-divided-over-robot-ai-artificial-intelligence-personhood/↩

CONCLUSIONI

COME PREVENIRE LA ARTIFICIAL IDIOCY?

Così come il dominio dell'AI è tutt'altro che "inevitabile", al tempo stesso occorre tuttavia attrezzarsi per tempo affinchè non si realizzi l'avvento della sua nemesi, ovvero la *Artificial Idiocy*.

In tal senso, occorre agire a vari livelli, piuttosto che perseguire in via prioritaria l'ipotesi *semplicistica* (e per molti versi *irrealistica*) di "regolamentare" la tecnologia.

Come abbiamo visto, infatti, l'intervento prematuro e mirato del regolatore pubblico può comportare più svantaggi che vantaggi, e rischia persino di ottenere l'effetto contrario rispetto a quello dichiarato e desiderato.

In realtà, più che regolamentare il settore, sarebbe necessario rinsaldare i principi normativi già esistenti e nel caso essi si dimostrino insufficienti, introdurre **regole generali e astratte** per garantire i diritti dei cittadini, piuttosto che misure che possano **condizionare o indirizzare** l'evoluzione della tecnologia verso obiettivi prefissati, che ad oggi sono difficili da definire e da valutare nella loro desiderabilità.

In altri termini, come spesso capita, non è la tecnologia il vero problema, quanto piuttosto la necessità da parte delle **istituzioni democratiche** di **non abdicare** al loro ruolo di preservare la garanzia del **rispetto dei diritti civili** e del loro corretto esercizio, cedendo alla tentazione di **delegare alla tecnologia** (rectius: alle aziende tecnologiche e ai tecnocrati) tale ruolo, alla luce di una presunta maggiore abilità e adeguatezza nel risolvere i problemi della contemporaneità, in virtù di supposte capacità

taumaturgiche che sempre più spesso vengono associate ad essa in maniera acritica e preconcetta.

In questo senso, è indispensabile che vengano coinvolti i vari *stakeholders*, tra cui vanno ricompresi i rappresentanti delle categorie svantaggiate, nelle diverse fasi di introduzione delle tecnologie *dirompenti* (disruptive) che sono suscettibili di determinare impatti rilevanti nella quotidianità dei cittadini (il riferimento alle self-driving cars è tutt'altro che casuale).

Coinvolgendo i diversi stakeholders è possibile anche affrontare i problemi derivanti dalla **ottimizzazione delle singole funzioni obiettivo**, consentendo così di ovviare alle "semplificazioni" algoritmiche, tenendo in debito conto le diverse implicazioni derivanti dalla **complessità** del contesto sociale, pervenendo di conseguenza a un miglior bilanciamento dei diversi interessi in gioco.

In ultima analisi, adottare sempre un **atteggiamento critico** nel valutare l'attendibilità delle notizie e delle informazioni che riceviamo, rimane tuttora il rimedio più valido per non cadere vittime degli **incantesimi** e della **fascinazione** che indubbiamente caratterizzano la tecnologia in generale, e la cosiddetta intelligenza artificiale in particolare.

In tal senso, è utile ricordare un noto consiglio originariamente suggerito da Enrico Fermi, secondo il quale **affermazioni eccezionali** e fuori dal comune che contrastano con le conoscenze consolidate, richiedono il supporto di **evidenze** altrettanto **eccezionali**.

E l'*onere della prova* ricade su chi avanza tali affermazioni eccezionali, non su chi le contesta: troppo spesso assistiamo invece alla diffusione di **affermazioni eclatanti** sugli straordinari risultati conseguiti dalla tecnologia, senza che vi siano **evidenze altrettanto straordinarie** a supporto di tali affermazioni.

Di fronte al legittimo scetticismo, i *tecno-sciovinisti* sono soliti ribattere richiedendo pretestuosamente agli scettici la dimostrazione della *falsità* delle loro affermazioni, quando invece sarebbe compito loro dimostrare l'attendibilità (quando non la *verità*) delle proprie affermazioni.

Detto in altri termini, per rimanere all'ambito della *Artificial Idiocy*, il fatto di non poter dimostrare l'*impossibilità* teorica che le macchine possano acquisire in futuro qualche forma di *consapevolezza*, non autorizza a darla per *possibile* o addirittura *inevitabile*.

Ma questi sono argomenti buoni per una trattazione dedicata e per un altro libro...

NOTE SULL'AUTORE

Alessandro Parisi è un professionista IT da oltre 20 anni, con una significativa esperienza come Computer Scientist, è specialista nei settori della Cybersecurity, Artificial Intelligence e Blockchain.

Ha maturato una vasta esperienza professionale in contesti organizzativi e decisionali caratterizzati da elevata complessità, supportando le aziende nella adozione delle tecnologie innovative come strumenti strategici per proteggere e valorizzare le risorse aziendali.

E' autore di pubblicazioni specialistiche, tra cui "Hands-on Artificial Intelligence for Cybersecurity", adottato come testo di riferimento da diverse Università internazionali, e del testo "Securing Blockchain Networks like Ethereum and Hyperledger Fabric".

Fin dal 2006 si occupa di Privacy Compliance, è autore del testo "Sicurezza Informatica e Tutela della Privacy", e da febbraio 2022 è stato designato come Data Protection Experts dal Council of Europe (CoE).

Attualmente ricopre l'incarico di Responsabile Ricerca e Sviluppo per conto del gruppo Meridian.

Per maggiori informazioni sul profilo professionale dell'Autore, è possibile visitare il link:
https://www.linkedin.com/in/parisialessandro/